European Journal of Operational Research 318 (2024) 911–926

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Analytics, Computational Intelligence and Information Management

The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective

Emanuele Borgonovo [a], Elmar Plischke [b], Giovanni Rabitti [c,*]

[a] Bocconi Institute for Data Science and Analytics and Department of Decision Sciences, Bocconi University, Milan, Italy
[b] Institute of Resource Ecology, Helmholtz-Zentrum Dresden-Rossendorf, Germany
[c] Department of Actuarial Mathematics and Statistics, Heriot-Watt University and Maxwell Institute for Mathematical Sciences, Edinburgh, UK

ARTICLE INFO

Keywords:
Analytics
Sensitivity analysis
Game theory
Aggregation of local importance effects
Interactions

ABSTRACT

Predictive models are increasingly used for managerial and operational decision-making. The use of complex machine learning algorithms, the growth in computing power, and the increase in data acquisitions have amplified the black-box effects in data science. Consequently, a growing body of literature is investigating methods for interpretability and explainability. We focus on methods based on Shapley values, which are gaining attention as measures of feature importance for explaining black-box predictions. Our analysis follows a hierarchy of value functions, and proves several theoretical properties that connect the indices at the alternative levels. We bridge the notions of totally monotone games and Shapley values, and introduce new interaction indices based on the Shapley-Owen values. The hierarchy evidences synergies that emerge when combining Shapley effects computed at different levels. We then propose a novel sensitivity analysis setting that combines the benefits of both local and global Shapley explanations, which we refer to as the "glocal" approach. We illustrate our integrated approach and discuss the managerial insights it provides in the context of a data-science problem related to health insurance policy-making.

1. Introduction

Machine learning (ML) tools are increasingly used in operations-research and the management sciences (ORMS) (De Bock et al., 2023). Decision-makers adopt recommendations from algorithmic sources in a variety of applications and sectors such as health-care (Naumzik et al., 2023; Ni et al., 2020), insurance (Florez-Lopez, 2007) and finance (Chen et al., 2024; Kriebel & Stitz, 2022; Nazemi et al., 2022).

While ML tools hold a promise of efficiency and of improving decision-making, they are associated with risks related to their opacity and lack of transparency. Lebovitz et al. (2022) report that health professionals in high-stake decisions do not trust machine learning advice when perceived as opaque. Fu et al. (2022) and Rudin (2019) outline the fairness problems associated with ML tools that are proprietary and not open-source. This creates a need for methods that aid the explanation of ML models decisions, as set forth in the XAIOR (Explainable Artificial Intelligence for Operational Research) framework of De Bock et al. (2023).

Methods that allow us to open the model black-box become crucial to increase trust in predictions. Murdoch et al. (2019) discuss the respective roles of interpretability and explainability. Interpretability is a modus operandi, in which the analyst favors a more transparent model over a more complex one when predictive accuracy is similar. Explainability entails generating post-hoc explanations through tools that inspect the input–output mapping after it has been trained. As underlined by Guidotti et al. (2018), model inspection involves visualizing marginal effects, determining feature importance, and identifying feature interactions. Recent literature has noted the overlap between explainable AI tasks and sensitivity analysis (Scholbeck et al., 2023). To illustrate, variance-based methods (Saltelli et al., 2000; Wagner, 1995), distribution-based techniques (Baucells & Borgonovo, 2013; Chatterjee, 2021; Wiesel, 2022), and Shapley values (Owen, 2014) are commonly used in both realms.

Shapley values have been introduced by Shapley (1953) as a mechanism to distribute the value of a game among its players. They have been intensively studied as an allocation method in Economics and they have found extensive applications also in ORMS. Studies like Balog et al. (2017), Bergantiños et al. (2023), Csóka et al. (2022), and Lindelauf et al. (2013), have applied Shapley values to identify key drivers in terrorist activity networks, in financial risk allocation, in liability negotiation for insolvent firms, and fleet allocation, respectively. They have also been intensively used to define feature importance measures in machine learning applications (Cohen et al., 2007; Štrumbelj &

[*] Corresponding author.
E-mail address: g.rabitti@hw.ac.uk (G. Rabitti).

https://doi.org/10.1016/j.ejor.2024.06.023
Received 13 December 2023; Accepted 14 June 2024
Available online 22 June 2024

Kononenko, 2010; Štrumbelj et al., 2009; Sundararajan & Najmi, 2019). In ORMS, the Shapley-additive explanations (SHAPs) of Lundberg and Lee (2017) are widely used as post-hoc explanations. Ahmed et al. (2024), Chen et al. (2024), Senoner et al. (2022), and Sobrie et al. (2023) employ them, respectively, in a data-driven analysis to improve process quality in semiconductor manufacturing, in the context of imbalanced credit scoring, in a data-driven approach for improving efficacy in railway operations, and in developing an integrated framework (from descriptive to prescriptive analytics) for predicting car accident severity.

One of the advantages of formulating feature importance measures through Shapley values is the flexibility in choosing the value function. This versatility allows us to obtain a variety of importance indices (Sundararajan & Najmi, 2019). However, it also leads to a fragmentation of the definitions and to the absence of a systematic approach for inferring managerial insights. To illustrate, choosing a value function that anchors the Shapley value to a given instance as in Štrumbelj et al. (2009) yields indices that explain individual predictions, while choosing a value function that considers the variance of the target as in Owen (2014) yields global indices that provide an overall (dataset-level) importance of a feature. However, what is the insight that the analyst finally looking for?

Our goal is to strengthen the link between Shapley values and sensitivity analysis for the extraction of modeling and managerial insights. To do so, we proceed as follows. We start connecting value functions hierarchically from a local to a global scale. At the local level, we introduce the notion of finite-change Shapley values and study them in connection with the decomposition of a finite change in the model predictions. We then set out a *glocal* approach that bridges the local and global scale to provide an increased understanding of the model's behavior across data. We finally consider global value functions that allow a full uncertainty quantification, focusing on the target variance as in Owen (2014) and on the target distribution as in Sarazin et al. (2020). We complete the framework with the recent proposals of generalized Shapley values (called Shapley-Owen values) (Dhamdhere et al., 2020; Rabitti & Borgonovo, 2019) for the quantification of joint contributions, and show that these measures for interactions can be directly obtained in parallel to the Shapley values in this common framework.

To illustrate how the sensitivity indices support the viewpoints of alternative stakeholders, we discuss a data-driven analysis of insurance premia, considering the perspective of an insured individual, a regulator, and an insurance company.

The remainder of the paper is organized as follows. Section 2 provides a concise literature review and introduces the Shapley value. Section 3 presents the hierarchical framework of the value functions adopted in the literature. Section 4 presents the finite-change Shapley values and describes their connection to partial derivatives as measures of importance. Section 5 presents the aggregation methods of partial derivatives producing global importance measures. Section 6 introduces the Shapley effects based on the variance decomposition. Section 7 introduces the generalized Shapley values for interaction quantification and extends previous results for individual Shapley values. Section 7.4 contains a discussion. Section 8 presents an application to a real dataset of medical insurance premiums. Section 9 concludes and provides future research directions.

In the Appendix in the Supplementary material, we provide the proofs, an introduction to the cohort estimation of Shapley effects, the Shapley chain rule for neural networks, and we extend the Shapley value framework to value functions that consider the whole output distribution instead of its variance.

2. The Shapley value

Consider a function $v : 2^N \to \mathbb{R}$, with $v(z) \in \mathbb{R}$ and $z \subseteq N$, where $N = \{1, 2, \dots, n\}$, such that $v(\emptyset) = 0$. This function is called value function and quantifies the value of a game played by the z players. We denote the Shapley value for the player i as Sh_i, where $i = 1, 2, \dots, n$. Shapley (1953) states the following axioms:

1. (Efficiency) $\sum_{i=1}^{n} Sh_i = v(N)$;
2. (Symmetry) If $v(z \cup i) = v(z \cup j)$ for every $z \subseteq N \setminus \{i, j\}$, then $Sh_i = Sh_j$;
3. (Additivity) Given two value functions v and v' with associated Shapley values Sh and Sh' respectively, then the game with the value function $v(z) + v'(z)$ has Shapley value $Sh_i + Sh'_i$ for all $i \in N$.

and shows that the following is the unique attribution method that satisfies them:

$$Sh_i = \frac{1}{n} \sum_{z \subseteq N \setminus \{i\}} \binom{n-1}{|z|}^{-1} (v(z \cup i) - v(z)). \tag{1}$$

Intuitively, the quantity Sh_i is a weighted mean of the differences in the value of the game caused by the participation of player i to all possible coalitions. Besides the three above mentioned properties, Shapley also shows that if $v(z \cup i) = v(z)$ for every $z \subseteq N$, then $Sh_i = 0$. A player satisfying this condition is called a null player.

The same quantity can be obtained through an alternative axiomatic characterization in which a central role is played by the Moebius transform of the value function (Rota, 1964):

$$m(v) = \sum_{u \subseteq v} (-1)^{|v|-|u|} v(u) \tag{2}$$

for any subset of inputs $v \subseteq N$. Note that $m(\emptyset) = 0$ and $m(\{i\}) = v(\{i\})$ for all $i = 1, \dots, n$. Using Eq. (2) it is possible rewrite the ith Shapley value as:

$$Sh_i = \sum_{z \subseteq N, i \in z} \frac{m(z)}{|z|}, \tag{3}$$

where m is the Moebius transform of v. When $m(v)$ is non-negative for every $v \subseteq N$, the game is said to be totally monotone (Owen, 2014). In Section 4.2 we connect the notion of totally monotone games to a form of monotonicity of the black-box model to be learned. This allows us to characterize the Shapley values when the analysts incorporate this functional knowledge on it.

3. A hierarchy of value functions for Shapley values in machine learning

In machine learning, analysts deal with the problem of quantifying or forecasting one or more quantities of interest, T, typically called target, as a function of one or more independent variables $\mathbf{X}$, typically called features. It is assumed that nature links $\mathbf{X}$ and T via a mathematical model and this model can be learned via data collection. We regard the machine learning model as an input–output function f that maps $\mathbf{X}$ and T through

$$T = f(\mathbf{X}), \tag{4}$$

where $f : \mathcal{X} \to \mathbb{R}$, with $\mathcal{X} \subseteq \mathbb{R}^n$. One also regards both features and targets as random variables on the reference measure space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$. We assume that $\mathbf{X} = (X_1, X_2, \dots, X_n)$ admits a probability distribution μ. We also assume that the function f is measurable and that the input space $\mathcal{X} \equiv \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ is the Cartesian product space of the individual input spaces $\mathcal{X}_i$.

The symbols t and $\mathbf{x}$ denote, respectively, a realization of T and $\mathbf{X}$. For instance, in a neural network, t might be the output of a target neuron of interest and $\mathbf{x} = (x_1, x_2, \dots, x_n)$ the set of neuron values in one or more preceding layers that are input to f (Shrikumar et al., 2017).

The application of Shapley values in a machine learning context is then based on two intuitions. The first is that features now represent players. The second is that one creates a value function v such that the corresponding Shapley value are consistent with the application at

Table 1
Main notation and symbols used in the paper.

Symbol	Name	Equation	Note
$v(z)$	Generic value function		
	baseline value, finite change	(5)	Local
	all-baseline value	(6)	Glocal
	(squared) cohort value	(7),(8)	Global
	regression value	(10)	Global
Sh_i	Generic Shapley value	(1),(3)	
$Sh_i^{\mathbf{x}^0\to\mathbf{x}}$	Finite-change Shapley value	(17)	Linked to (19)
D_i	Differential importance index	(26)	Linked to (28)
ζ_i	Derivative-based global sensitivity measure	(32)	Proposition 10
Sh_i^{VB}	(Variance-based) Shapley effects	(43)	Theorem 11
Sh_i^{SC}	Squared Cohort Shapley value		Theorem 12
Sh_s^{v}	Generic Shapley-Owen value	(48)	
ShT_s	Shapley-Taylor interaction index	(53)	
$Sh_s^{\mathbf{x}^0\to\mathbf{x}}$	Finite-change Shapley-Owen value	(51)	Theorem 16
Sh_s^{VB}	Variance-based Shapley-Owen value	(56)	Theorem 17
Sh_s^{SC}	Squared-cohort Shapley-Owen value		Theorem 18

hand. In the literature, a variety of formulations have been proposed. These definitions are scattered all over the literature as they are often application-dependent, and a unifying framework is missing. To propose a unifying view, we start with local formulations and move up to global formulations (Table 1).

We start with the local formulation. We fix a base case value $\mathbf{x}^0$ in the feature space and denote with $t^0 = f(\mathbf{x}^0)$ the corresponding model prediction. We then consider an alternative point $\mathbf{x}$, called sensitivity case, and the corresponding prediction $t = f(\mathbf{x})$. Several works then consider the problem of explaining the difference $\Delta t = t - t^0 = f(\mathbf{x}) - f(\mathbf{x}^0)$ (Shrikumar et al., 2017; Sundararajan & Najmi, 2019). A natural choice to define Shapley effects in this context is to rely on the so-called baseline value function

$$v(z) = f(\mathbf{x}_z : \mathbf{x}^0_{-z}) - f(\mathbf{x}^0), \tag{5}$$

where $(\mathbf{x}_z : \mathbf{x}^0_{-z})$ denotes the point obtained by: (a) shifting the features with indices in z to the sensitivity case and (b) by keeping the remaining features at the base case. This value function has been used in the sensitivity analysis of neural networks in Sundararajan and Najmi (2019) and in the DeepLIFT explanation model (Shrikumar et al., 2017). Because the value function (5) produces Shapley values which depend on the initial evaluation point $\mathbf{x}^0$, Mase et al. (2020) and Sundararajan and Najmi (2019) call the resulting indices *baseline Shapley values*.

The value function (5) does not take feature uncertainty into consideration. Under uncertainty, Štrumbelj and Kononenko (2010) define the value function

$$v(z) = \mathbb{E}\left[f(\mathbf{x}_z : \mathbf{X}^0_{-z})\right] - \mathbb{E}\left[f(\mathbf{X}^0)\right], \tag{6}$$

where the point $\mathbf{x}$ (and its projections $\mathbf{x}_z$) are kept constant. Note that the value function (6) is the average of the value function (5) over initial points. Štrumbelj and Kononenko (2010) use it to quantify feature importance in classification problems (see also Janzing et al., 2019; Mase et al., 2020). Mase et al. (2020) call the Shapley value generated by the value function (6) *all-baseline Shapley value*. The intuition is that the average of the model output constitutes the reference initial point (the baseline). We also note that the first summand in Eq. (6) is the so-called partial dependence function $h_z(x_z) = \mathbb{E}\left[f(\mathbf{x}_z : \mathbf{X}^0_{-z})\right]$ of Friedman and Popescu (2008) (see also Hooker, 2004). This partial dependence function is a marginal expectation with respect to a subset of inputs and is commonly used to visualize the marginal relationship between the output and the inputs in a reduced feature space (Goldstein et al., 2015; Guidotti et al., 2018).

Alternatively, Štrumbelj and Kononenko (2014) consider the value function

$$v(z) = \mathbb{E}\left[f(\mathbf{X})|\mathbf{X}_z = \mathbf{x}_z\right] - \mathbb{E}\left[f(\mathbf{X})\right], \tag{7}$$

where $\mathbb{E}\left[f(\mathbf{X})|\mathbf{X}_z = \mathbf{x}_z\right]$ is the conditional nonparametric regression curve. Shapley values based on (7) are used for explaining individual feature contribution in black-box predictive models in several works (Aas et al., 2019; Datta et al., 2016; Lundberg et al., 2020; Lundberg & Lee, 2017; Molnar, 2018; Štrumbelj & Kononenko, 2014; Sundararajan & Najmi, 2019). The SHAP method of Lundberg and Lee (2017) is based on partial dependence functions, and thus on the value function in (6).[1] Differently, starting from the conditional regression curve in Eq. (7), Mase et al. (2020) take a data-driven approach. They call Eq. (7) the *cohort* value function. When features are independent, (6) and (7) coincide.

Mase et al. (2020) introduce the squared version of the value function in Eq. (7) as

$$v(z) = \left(\mathbb{E}\left[f(\mathbf{X})|\mathbf{X}_z = \mathbf{x}_z\right] - \mathbb{E}\left[f(\mathbf{X})\right]\right)^2, \tag{8}$$

which they call the *squared cohort* value function. Owen (2014) introduces the value function

$$v(z) = \mathbb{V}\left[\mathbb{E}\left[f(\mathbf{X})|\mathbf{X}_z\right]\right] \tag{9}$$

to obtain global sensitivity measures in the context of the sensitivity analysis of computer simulators. The value function in Eq. (9) allows one to define variance-based sensitivity measures for models with dependent features (Benoumechiara & Elie-Dit-Cosaque, 2019; Iooss & Prieur, 2019; Owen & Prieur, 2017; Song et al., 2016). One may observe that the expected value of the squared cohort value function (8) yields (9). Similarly, the value function in Eq. (9) can be obtained taking the variance of the value function in Eq. (7). Thus, there is a strict link between the value functions (7), (8) and (9). The Shapley values constructed from these value functions are feature importance measure on a global scale, because the uncertainty in $\mathbf{X}$ is taken into account.

There is however an important point to raise. Differently from (9), the value functions (7) and (8) depend on the specific conditioning value $\mathbf{X} = \mathbf{x}$. This makes the sensitivity scale of (6), (7) and (8) "glocal". This glocal approach strikes a balance between local and global explanations while still focusing on a single observation, making it a useful tool to investigate the drivers of the model output given a target value of $\mathbf{x}$. By identifying the most important features for a specific instance's prediction, the glocal approach can lead to local insights that would otherwise be lost. To illustrate, assume that the model is an insurance pricing model used to assign policyholders a premium based on their specific risk profiles. With reference to a target

[1] See the *Reading SHAP values from partial dependence plots* Section at https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An_introduction_to_explainable_AI_with_Shapley_values.html, accessed on May 30 2024.

policyholder, the most important features measured by the Shapley values obtained from (7) and (8) can differ from the Shapley values obtained for the whole portfolio using (9). We will illustrate this point in our case study (Section 8).

When the machine learning problem involves linear regression models, Lipovetsky and Conklin (2001) introduce the Shapley regression values for feature importance. Shapley regression values are studied in Grömping (2007, 2015) and Huettner and Sunder (2012). The value function in this case is

$$v(z) = R_z^2, \tag{10}$$

where R_z^2 is the well-known goodness-of-fit measure when only the features in z are included in the model. Eq. (10) coincides with the normalized version of Eq. (9) in the case of a linear model with independent inputs (Saltelli et al., 2000), that is

$$R_z^2 = \frac{\mathbb{V}\left[\mathbb{E}\left[f(\mathbf{X})|\mathbf{X}_z\right]\right]}{\mathbb{V}\left[f(\mathbf{X})\right]}.$$

Sarazin et al. (2020) have recently considered Shapley effects associated with the value function generated by a moment independent global sensitivity measure. We address this value function in greater detail in Appendix 4.

This discussion offers a hierarchical path through value functions used to formulate Shapley values in machine learning. These formulations allow us not only to investigate the relationships between Shapley indices formulated using alternative value functions, but also the relationships between Shapley indices and other sensitivity measures used in machine learning as well as computer simulation literature.

4. Finite-change sensitivity analysis for local model explanations

This section focuses on the Shapley formulated using the finite-change (or baseline) value function in Eq. (5). In the first part, we connect the sensitivity measure of Lundberg and Lee (2017), Shrikumar et al. (2017) to the forward-difference decomposition of a finite change in the model predictions. In the second part, we outline a bracketing property for Shapley values that holds when the input–output mapping satisfies a given monotonicity requirement. In the third part, we discuss the limiting behavior of the Finite Change Shapley value and demonstrate the link to differential importance measures.

4.1. Shapley values for Finite Changes

We consider the finite change $\Delta t = t - t^0 = f(\mathbf{x}) - f(\mathbf{x}^0)$ which represents the difference in the target registered when the features vary from the base case to the sensitivity case. Shrikumar et al. (2017) define the contribution score $C_{\Delta x_i \Delta t}$ of $\Delta x_i = x_i - x_i^0$ as *the amount of difference-from-reference in t that is attributed [...] to the difference from-reference in x_i* (p. 3). They impose the condition

$$\Delta t = \sum_{i=1}^{n} C_{\Delta x_i \Delta t}. \tag{11}$$

Now, it is known that one can expand Δt through finite differences as Borgonovo (2010), Kuo et al. (2010) and Rabitz and Alis (1999)

$$\Delta t = f(\mathbf{x}) - f(\mathbf{x}^0) = \sum_{i=1}^{n} \phi_i^{\mathbf{x}^0\to\mathbf{x}} + \sum_{i<j} \phi_{i,j}^{\mathbf{x}^0\to\mathbf{x}} + \cdots + \phi_{1,2,\ldots,n}^{\mathbf{x}^0\to\mathbf{x}}. \tag{12}$$

In Eq. (12) the $2^n - 1$ finite-change terms are found from

$$\begin{cases} \phi_i^{\mathbf{x}^0\to\mathbf{x}} = f(x_i : \mathbf{x}_{-i}^0) - f(\mathbf{x}^0), \\ \phi_{i,j}^{\mathbf{x}^0\to\mathbf{x}} = f(x_{i,j} : \mathbf{x}_{-i,j}^0) - \phi_i^{\mathbf{x}^0\to\mathbf{x}} - \phi_j^{\mathbf{x}^0\to\mathbf{x}} - f(\mathbf{x}^0), \\ \ldots \end{cases} \tag{13}$$

The first order terms $\phi_i^{\mathbf{x}^0\to\mathbf{x}}$, $i = 1, 2, \ldots, n$, are called main effects and the higher-order terms are called interaction effects. The superscript $\mathbf{x}^0 \to \mathbf{x}$ makes explicit that every effect is referred to the finite change from the reference point $\mathbf{x}^0$ to the sensitivity case $\mathbf{x}$. Given the decomposition in Eq. (13), it is possible to define the total finite-change effect of feature x_i as Borgonovo (2010) and Borgonovo and Rabitti (2023)

$${}^{T}\phi_i^{\mathbf{x}^0\to\mathbf{x}} = \phi_i^{\mathbf{x}^0\to\mathbf{x}} + \sum_{k=2}^{n} \sum_{|z|=k, i\in z} \phi_z^{\mathbf{x}^0\to\mathbf{x}}, \tag{14}$$

which includes all the effect terms to which x_i contributes to. Thus, ${}^{T}\phi_i^{\mathbf{x}^0\to\mathbf{x}}$ is a measure of the total impact of x_i to the target change Δt. Similarly, we can define the overall interaction effect associated with x_i as the difference between the total and the main effects of x_i:

$${}^{I}\phi_i^{\mathbf{x}^0\to\mathbf{x}} = {}^{T}\phi_i^{\mathbf{x}^0\to\mathbf{x}} - \phi_i^{\mathbf{x}^0\to\mathbf{x}}. \tag{15}$$

Analogously to the global sensitivity analysis setting (Owen, 2014), we define the lower and upper sensitivity indices of z for the finite change from $\mathbf{x}^0$ to $\mathbf{x}$ as

$$\underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}} = \sum_{u\subseteq z} \phi_u^{\mathbf{x}^0\to\mathbf{x}} \quad \text{and} \quad \overline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}} = \sum_{u: u\cap z\neq\emptyset} \phi_u^{\mathbf{x}^0\to\mathbf{x}}. \tag{16}$$

These indices allow analysts to understand the role of the zth features in determining the change Δt. The upper index $\overline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$ is a measure of the total effect of the feature group: if it is close to zero then the zth features can be considered uninfluential. Note that if $z = \{i\}$ then $\overline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$ equals ${}^{T}\phi_i^{\mathbf{x}^0\to\mathbf{x}}$. Instead, a large absolute value of the lower index $\underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$ indicates that the group of features contribute significantly to the change Δt. We can then link these finite-change sensitivity indices with the Shapley value.

Definition 1. For the finite change $\mathbf{x}^0 \to \mathbf{x}$ we call the Shapley value with value function $v(z) = \underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$ the *finite-change Shapley value* of the feature i.

We denote the finite-change Shapley value with $Sh_i^{\mathbf{x}^0\to\mathbf{x}}$. It holds

$$\begin{aligned} Sh_i^{\mathbf{x}^0\to\mathbf{x}} &= \frac{1}{n} \sum_{z\subseteq[n]\backslash i} \binom{n-1}{|z|}^{-1} \left(\underline{\tau}_{z\cup i}^{\mathbf{x}^0\to\mathbf{x}} - \underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}\right) \\ &= \frac{1}{n} \sum_{z\subseteq[n]\backslash i} \binom{n-1}{|z|}^{-1} \sum_{v\subseteq z} \phi_{v\cup i}^{\mathbf{x}^0\to\mathbf{x}}, \end{aligned} \tag{17}$$

where the terms $\phi_{v\cup i}^{\mathbf{x}^0\to\mathbf{x}}$ are the finite changes in Eq. (13). In addition, we have the following characterization.

Theorem 2. *Given the value function $v(z) = \underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$, the finite-change Shapley value of X_i is*

$$Sh_i^{\mathbf{x}^0\to\mathbf{x}} = \phi_i^{\mathbf{x}^0\to\mathbf{x}} + \sum_{k=2}^{n} \frac{1}{k} \sum_{|z|=k, i\in z} \phi_z^{\mathbf{x}^0\to\mathbf{x}}. \tag{18}$$

All proofs are in Appendix 1. The Shapley value $Sh_i^{\mathbf{x}^0\to\mathbf{x}}$ in Eq. (18) and the finite-change total effect ${}^{T}\phi_i^{\mathbf{x}^0\to\mathbf{x}}$ of the feature x_i in (14) contain the same finite changes. Nonetheless, the sum of $Sh_i^{\mathbf{x}^0\to\mathbf{x}}$ equals to Δt. The finite-change Shapley effect $Sh_i^{\mathbf{x}^0\to\mathbf{x}}$ can be connected to the contribution score used in Lundberg and Lee (2017) and Shrikumar et al. (2017) as follows.

Proposition 3. *$C_{\Delta x_i \Delta t}$ is a finite change Shapley value, i.e.*

$$C_{\Delta x_i \Delta t} = Sh_i^{\mathbf{x}^0\to\mathbf{x}}. \tag{19}$$

Next, we establish the equivalence of finite-change and baseline Shapley values.

Theorem 4. *Consider two points $\mathbf{x}^0$ and $\mathbf{x}$. Then, the baseline value function $v^{BS}(z) = f(\mathbf{x}_z : \mathbf{x}_{-z}^0) - f(\mathbf{x}^0)$ and the finite-change value function $v^{FC}(z) = \underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$ coincide.*

An important consequence of Theorem 4 is that the Shapley values for the baseline and the finite-change value functions coincide. By Theorem 2, the expression of these Shapley values is given by Eq. (18). Mase et al. (2020) already proved this characterization for the baseline Shapley values but only in the specific case of a function f defined on $\{0,1\}^n$ with reference point $\mathbf{x}^0 = (0,0,\ldots,0)$ and the sensitivity point $\mathbf{x} = (1,1,\ldots,1)$. Thus, Theorem 2 extends the result of Mase et al. (2020) showing that this characterization holds in general.

4.2. The case of monotonic models

We now explore the situation in which the response variable is required to be monotonic with respect to some input variables. Authors include monotonic constraints on the model for interpretability reasons (Rudin et al., 2022) and to increase the predictive accuracy (Dugas et al., 2009). For instance, several neural network architectures have been developed for totally and partially monotonic relationships (Daniels & Velikova, 2010; Dugas et al., 2009; Sill, 1998). Sundararajan and Najmi (2019) prove that, if the model is monotone in one feature, then the Shapley value of this feature increases as this variable increases. In this section, we focus our attention on a particularly useful notion of monotonicity due to Rüschendorf (2013).

Definition 5. A multivariate function f is said to be Δ-monotone if $\phi_z^{\mathbf{x}^0\to\mathbf{x}} \geq 0$ whenever $x_i \geq x_i^0$ for all $z \subseteq N$.

This notion of Δ-monotonicity is stronger than monotonicity, because it implies that $\Delta t \geq 0$ and all the finite-changes are positive. Using this notion, we can show that the following inequality holds for finite-change Shapley values.

Proposition 6. *Assume that the input–output map f is Δ-monotone. Then*

$$0 \leq \phi_i^{\mathbf{x}^0\to\mathbf{x}} \leq Sh_i^{\mathbf{x}^0\to\mathbf{x}} \leq {}^T\phi_i^{\mathbf{x}^0\to\mathbf{x}}. \tag{20}$$

We call (20) bracketing property in analogy with Owen (2014), where it is proven for the value function in Eq. (9) (see (44) in Section 6).

4.3. Connection to differential sensitivity measures

We now characterize the Shapley value (18) in terms of partial derivatives. Shrikumar et al. (2017) define the multiplier $M_{\Delta x \Delta t}$ as the contribution of the change Δx to the change Δt divided by Δx, that is

$$M_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}, \tag{21}$$

which by (19) becomes

$$M_{\Delta x \Delta t} = \frac{Sh^{\mathbf{x}^0\to\mathbf{x}}}{\Delta x}. \tag{22}$$

This index is analogous to a partial derivative which leads us to the following observations, and indeed several approaches to the intepretability of neural networks are based on partial derivatives (Baehrens et al., 2010; Kowalski & Kusy, 2018b; Montavon et al., 2017; Yeung et al., 2010). In particular, Engelbrecht et al. (1995), Zurada et al. (1994) define the input–output sensitivity of the trained output as

$$S_i^{(0)} = \frac{\partial f\left(\mathbf{x}^0\right)}{\partial x_i}. \tag{23}$$

We now prove the connection between Shapley values and partial derivatives at the infinitesimal scale.

Proposition 7. *Assume that the model f is differentiable. As $\mathbf{x} \to \mathbf{x}^0$ then $M_{\Delta x_i \Delta t} \to \frac{\partial f(\mathbf{x}^0)}{\partial x_i}$ for all $i = 1, 2, \ldots, n$.*

Proposition 7 shows the relationship between the finite-change Shapley value of variable i and the corresponding partial derivative. Note that, when the model admits positive first-order partial derivatives (Dugas et al., 2009), the finite-change Shapley values are positive.

Consider now that the features are denominated in different units. Then, (23) cannot be used to rank them, as one is comparing heterogeneous quantities. However, one can recombine the gradient information to construct importance measures. For instance, for the sensitivity analysis of neural networks, Tsaih (1999) considers the importance measure

$$D_i^* = \frac{\frac{\partial f(\mathbf{x}^0)}{\partial x_i}}{\sum_{k=1}^n \left|\frac{\partial f(\mathbf{x}^0)}{\partial x_k}\right|}. \tag{24}$$

On the other hand, Horel et al. (2018) consider as a measure of local feature importance

$$D_i^{**} = \frac{\left(\frac{\partial f(\mathbf{x}^0)}{\partial x_i}\right)^2}{\sum_{k=1}^n \left(\frac{\partial f(\mathbf{x}^0)}{\partial x_k}\right)^2}. \tag{25}$$

Note that both (24) and (25) resemble the recombination of local sensitivities in the differential importance measure given by Borgonovo and Apostolakis (2001):

$$D_i = \frac{\frac{\partial f(\mathbf{x}^0)}{\partial x_i}\Delta x_i}{\sum_{k=1}^n \frac{\partial f(\mathbf{x}^0)}{\partial x_k}\Delta x_k}, \tag{26}$$

but with a caveat. In the proof of Proposition 7 (see Eq. (D.4) in Appendix 1) we have shown the approximation

$$Sh_i^{\mathbf{x}^0\to\mathbf{x}} \approx \frac{\partial f\left(\mathbf{x}^0\right)}{\partial x_i}\Delta x_i, \tag{27}$$

which connects Shapley value to differential importance indices. We have shown that, when $\mathbf{x} \to \mathbf{x}^0$, the relevance scores for the local sensitivity analysis of deep neural networks of Montavon et al. (2018) are Shapley values.

Proposition 8. *The differential importance measure D_i locally behaves as*

$$D_i \approx \frac{Sh_i^{\mathbf{x}^0\to\mathbf{x}}}{\sum_{k=1}^n Sh_k^{\mathbf{x}^0\to\mathbf{x}}} = \frac{Sh_i^{\mathbf{x}^0\to\mathbf{x}}}{\Delta t} \tag{28}$$

as $\mathbf{x} \to \mathbf{x}^0$, or equivalently

$$Sh_i^{\mathbf{x}^0\to\mathbf{x}} \approx D_i \cdot \Delta t. \tag{29}$$

When the features are expressed in the same units, we can assume that the finite changes are uniform (i.e., $\Delta x_i = \Delta x_j$) and hence the differential importance measure (26) becomes

$$D_i = \frac{\frac{\partial f\left(\mathbf{x}^0\right)}{\partial x_i}}{\sum_{k=1}^n \frac{\partial f\left(\mathbf{x}^0\right)}{\partial x_k}}. \tag{30}$$

We conclude with a cursory observation on differentiation-based indices. Gradients are one of the main sensitivity analysis tools used in the stochastic simulation. When the model output is stochastic, it is natural to assess the parametric sensitivity of the output response via sensitivity indices of the type $\frac{\partial \mathbb{E}_{\mathbf{X}}[f(\mathbf{X};\theta)]}{\partial \theta}$, where θ represents a parameter of interest. The problem of correctly calculating these types of sensitivities goes under the name of perturbation analysis (Glasserman, 1990). Applications range from risk assessment to finance, and can be found in works such as Hong and Liu (2009, 2010), Pesenti et al. (2021) and Tsanakas and Millossovich (2016).

5. From the infinitesimal to the global scale

In this section, we address those methods that are a bridge between local and global approaches. We observe that they have been independently developed in different research streams in the machine learning and design of computer experiments communities. These methods are based on the aggregation of local sensitivities computed at randomized locations in the input space.

We start with the works in computer experiments. We recall first the average importance index of Becker et al. (2018) and Campolongo et al. (2007)

$$\mu_i^* = \mathbb{E}\left[\left|\frac{\partial f(\mathbf{X})}{\partial X_i}\right|\right]. \tag{31}$$

The absolute value ensures that there are no cancellation effects when aggregating partial derivatives. Another relevant index based on this approach is the derivative-based global sensitivity measure

$$\zeta_i = \mathbb{E}\left[\left(\frac{\partial f(\mathbf{X})}{\partial X_i}\right)^2\right] \tag{32}$$

introduced in Sobol' and Kucherenko (2009). In the machine learning literature, we find the same sensitivity indicators in Wang et al. (2008) who use ζ_i in Eq. (32) for identifying feature importance in neural networks. Consider a set of points $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^K$ sampled from the probability distribution μ. Denote by $S_i^{(k)}$ the partial derivative (23) evaluated at the point $\mathbf{x}^k$ for $k = 1, 2, \dots, K$. Then, Engelbrecht et al. (1995), Kowalski and Kusy (2018a, 2018b) and Zurada et al. (1994) define for feature i

- the mean square average sensitivity index S_i^{MSA} as

$$S_i^{MSA} = \sqrt{\frac{\sum_{k=1}^K \left(S_i^{(k)}\right)^2}{K}}; \tag{33}$$

- the absolute value average sensitivity index S_i^{AVA} as

$$S_i^{AVA} = \frac{\sum_{k=1}^K \left|S_i^{(k)}\right|}{K}; \tag{34}$$

- the maximum sensitivity index S_i^{MAX} as

$$S_i^{MAX} = \max_{k=1,2,\dots,K} S_i^{(k)}. \tag{35}$$

We can then establish the following relationships between the two literature streams.

Proposition 9. *The absolute value sensitivity index S_i^{AVA} is an estimate of the average importance index $\mu_i^\star$ in Eq. (31). Moreover, the mean square average sensitivity index S_i^{MSA} is the square root of the estimate of the derivative-based global sensitivity measure in Eq. (32), that is*

$$S_i^{MSA} \approx \sqrt{\hat{\zeta}_i}.$$

Assume now that the analyst has performed a randomized evaluation of finite-changes Shapley values. Note that, in the same spirit, we can aggregate finite-change Shapley values computed at randomized locations in the input space.

Proposition 10. *The derivative-based global sensitivity measure in Eq. (32) can be approximated using replicated Shapley values as*

$$\zeta_i \approx \mathbb{E}\left[\left(\frac{Sh_i^{\mathbf{X}\to\mathbf{X}+\Delta\mathbf{X}}}{\Delta X_i}\right)^2\right] \tag{36}$$

for small values of Δ. Analogous approximations hold for the sensitivity indices S_i^{MSA}, S_i^{AVA} and S_i^{MAX}.

Propositions 9 and 10 suggest that it is possible to estimate the differential importance measure D_i in Eqs. (28), all global sensitivity measures in Eqs. (31)–(35), as well as approximated finite-change Shapley values at K randomized locations. In our application, we suggest to exploit the information yielded by the knowledge of Shapley values at several points before performing the aggregation. In this way, we gain additional insights at no further cost.

6. Shapley effects and variance-based global sensitivity measures

A further way for obtaining global importance measures consists in considering Eq. (9) as the value function (Owen, 2014), and is attracting increasing attention in the uncertainty quantification literature (Owen, 2014; Owen & Prieur, 2017; Plischke et al., 2021; Song et al., 2016). In this context Shapley values are called Shapley effects, following Song et al. (2016).

Assume for the moment that features are independent and that f is square integrable. Then, f can be written as Efron and Stein (1981) and Hoeffding (1948)

$$f(\mathbf{x}) = \sum_{z\subseteq N} f_z(\mathbf{x}_z), \tag{37}$$

where $\mathbf{x}_z$ are the components of $\mathbf{x}$ indexed by $z \subseteq N$ and the functions f_z are recursively defined by

$$f_z(\mathbf{x}_z) = \int \left(f(\mathbf{x}) - \sum_{l\subsetneq z} f_l(\mathbf{x}_l)\right) d\mu(\mathbf{x}_{-z}).$$

Eq. (37) is called classical functional ANOVA expansion of f. Component functions $f_z(\mathbf{x}_z)$ represent the effects of the features $\mathbf{x}_z$. The component functions $f_i(x_i)$, $i = 1, 2, \dots, n$, coincide with the partial dependence functions (Friedman & Popescu, 2008; Hooker, 2004). The ANOVA functional components are connected to the finite-change terms in (13) as

$$f_z(\mathbf{x}_z) = \int \phi_z^{\mathbf{x}^0\to\mathbf{x}} d\mu(\mathbf{x}^0), \tag{38}$$

where the integral is taken with respect to the initial point (also called anchor point in Kuo et al., 2010; Mase et al., 2020; Rabitz & Alis, 1999).

Letting $\sigma_z^2 = \mathbb{V}[f_z(X_z)]$, the orthogonality of the component functions allows the decomposition of the variance of f into $2^n - 1$ orthogonal terms (Efron & Stein, 1981):

$$\sigma^2 = \mathbb{V}[f(\mathbf{X})] = \sum_{z\subseteq N\setminus\{\emptyset\}} \sigma_z^2. \tag{39}$$

By Eq. (38), one finds that a link between the variance-components and the finite-change indices from Rabitz and Alis (1999):

$$\sigma_z^2 = \mathbb{V}\left[\mathbb{E}\left[\phi_z^{\mathbf{X}^0\to\mathbf{X}}\right]\right], \tag{40}$$

where the external variance is taken with respect to the final point $\mathbf{X}$ and the internal expectation with respect to the initial (anchor) point in the finite-change decomposition $\mathbf{X}^0$.

Natural sensitivity measures are then Sobol' indices σ_z^2 (Sobol', 1993). The two importance indices for subset z constructed from Sobol' indices are

$$\underline{\tau}_z^2 = \mathbb{V}\left[\mathbb{E}\left[f(\mathbf{X}) \mid \mathbf{X}_z\right]\right] = \sum_{l\subseteq z} \sigma_l^2 \tag{41}$$

and

$$\overline{\tau}_z^2 = \mathbb{E}\left[\mathbb{V}\left[f(\mathbf{X}) \mid \mathbf{X}_{-z}\right]\right] = \sum_{l:l\cap z\neq\emptyset} \sigma_l^2. \tag{42}$$

In particular, the index $\underline{\tau}_z^2$ represents the variance explained by $\mathbf{x}_z$ and can be considered as a natural importance measure for the subset of features with indices in z. Conversely, $\overline{\tau}_z^2$ is usually called the total effect of group z (Homma & Saltelli, 1996) and it can be interpreted as the expected remaining variance once the variables $\mathbf{X}_{-z}$ are known[2]. These indices satisfy $\underline{\tau}_z^2 \leq \overline{\tau}_z^2$ and $\overline{\tau}_z^2 = \sigma^2 - \underline{\tau}_{-z}^2$. These indices have been applied to the sensitivity analysis of neural networks in Cheng et al. (2019), Fernández-Navarro et al. (2017), Fock (2014), Kowalski and Kusy (2018a, 2018b) and Li and Chen (2018).

Owen (2014) proposes the variance-based value function $v(z) = \underline{\tau}_z^2$ and proves that, under feature independence, the Moebius transform of the value function $m(z)$ coincides with the Sobol' index σ_z^2 for any z. In a subsequent work, Song et al. (2016) prove that the Shapley effects using value functions (41) or (42) coincide.

Hence, the representation formula (3) under feature independence becomes

$$Sh_i^{VB} = \sum_{z \subseteq N, j \in z} \frac{\sigma_z^2}{|z|}. \tag{43}$$

Owen (2014) proves the following inequality

$$0 \leq \underline{\tau}_i^2 \leq Sh_i^{VB} \leq \overline{\tau}_i^2, \tag{44}$$

and calls it the bracketing property. It states that the Shapley effect of the ith feature is always comprised between its individual variance-based sensitivity measure and its total order index. This signals a parallelism to finite-change Shapley values for Δ-monotone functions (see Proposition 6). When features are independent, Shapley effects can be explicitly represented in terms of finite-change sensitivity indices.

Theorem 11. *Assume that f is L^2-integrable and that the features are independent. Then, we can write*

$$Sh_i^{VB} = \mathbb{V}\left[\mathbb{E}\left[\sum_{z \subseteq N, j \in z} \frac{\phi_z^{\mathbf{X}^0 \to \mathbf{X}}}{\sqrt{|z|}}\right]\right]. \tag{45}$$

This result has an operational implication. Suppose one has available sensitivity indices $\{\phi_z^{\mathbf{x}^\mathbf{k} \to \mathbf{x}^{\mathbf{k+1}}}\}_{k=0}^{K}$ calculated at $K+1$ randomized locations. These indices can be aggregated in two ways: via Eq. (17) to find $Sh_i^{\mathbf{x}^\mathbf{k} \to \mathbf{x}^{\mathbf{k+1}}}$, or via Eq. (45) to find the variance-based Shapley effects. Eq. (45) shows that Sh_i^{VB} provides information at a global scale but it based on information at a local scale. At a global scale, we obtain indication about the relative importance of the features. On a local scale, knowledge of ϕ_i at randomize location provides additional information about the local behavior of the model through the input space. Thus, insights at the local and global scale can be simultaneously obtained from the same machine learning model evaluations. Information on local patterns can be particularly useful to identify non-regular behavior of the model in localized regions of the input space, especially when some features can have only local and not global importance.

It is possible to find an alternative characterization of Sh_i^{VB} relaxing the input variables independence assumption in Theorem 11. Denote by $v^{SC}(z)$ the squared-cohort value function in Eq. (8). As written in Section 3, taking the expectation we find that

$$v^{VB}(z) = \mathbb{E}\left[v_{\mathbf{X}}^{SC}(z)\right], \tag{46}$$

where $v^{VB}(z)$ is the variance-based value function in Eq. (9). Then, denote by $Sh_i^{SC}(\mathbf{x})$ the Shapley value of the ith feature generated by the squared cohort value function $v_{\mathbf{x}}^{SC}(z)$. The symbol $Sh_i^{SC}(\mathbf{x})$ evidences the dependence on the point $\mathbf{x}$ on which we condition in Eqs. (7) and (8). By the additivity property of the Shapley value, the following holds.

Theorem 12 (*Mase et al., 2020*)**.** *Assume that f is L^2-integrable. Then, for the ith feature one finds*

$$Sh_i^{VB} = \mathbb{E}\left[Sh_i^{SC}(\mathbf{X})\right], \tag{47}$$

where the expectation is taken with respect to $\mathbf{X}$.

Eq. (47) shows that Sh_i^{VB} is, in fact, the aggregation of $Sh_i^{SC}(\mathbf{x})$ at randomized locations. Then displaying $Sh_i^{SC}(\mathbf{x})$ at the locations before the aggregation provides insights on the local model behavior. In particular, this allows one to understand whether the importance of a feature is regionally localized. Displaying this information is then a nice bridge between local and global sensitivity. We illustrate this point in our application, with a new graphical visualization of Shapley effects. We extend the result of Theorem 12 to the generalized Shapley values for interactions in Section 7.

7. Generalized shapley values for interaction quantification

This section is divided into three parts. Section 7.1 presents the Shapley-Owen value for interactions. Section 7.2 discusses the computation of the Shapley-Owen index via randomization of high-order finite differences. Section 7.3 presents variance-based Shapley-Owen effects and new results about their estimation.

7.1. The shapley-owen value

The increasing attention in individual feature importance parallels the interest for determining the relevance of feature interactions (Dhamdhere et al., 2020; Plischke et al., 2021; Rabitti & Borgonovo, 2019). Interaction quantification is a fundamental issue in explaining predictions of black-box machine learning models. However, classical Shapley values constitute an attribution method for individual features. To address this issue, Owen (1972) defines the Shapley value of the coalition of two features and Grabisch and Roubens (1999) extend this notion to any coalition size. The resulting index is called the Shapley-Owen value. This index represents the residual interaction value of a coalition of features $s \subseteq N$ whose components are indexed by $\{i_1, i_2, \ldots, i_s\}$. The intuition is that computing the value of a coalition helps to understand whether the coalition is producing more/less value than the sum of individual feature contributions.

The Shapley-Owen value for a coalition s with value function v is denoted by Sh_s^v and is defined by:

$$Sh_s^v = \sum_{u \subseteq N \setminus s} \frac{(n-|u|-|s|)!|u|!}{(n-|s|+1)!} \sum_{l \subseteq s} (-1)^{|s|-|l|} v(l \cup u). \tag{48}$$

For instance, consider this interaction index between the features of interest $s = \{i, j\}$. Eq. (48) becomes

$$Sh_{i,j}^v = \sum_{u \subseteq N \setminus \{i,j\}} \frac{(n-|u|-2)!|u|!}{(n-1)!} \left[v(u \cup \{i,j\}) - v(u \cup \{i\}) - v(u \cup \{j\}) + v(u)\right]. \tag{49}$$

The term $v(u \cup \{i,j\}) - v(u \cup \{i\}) - v(u \cup \{j\}) + v(u)$ in Eq. (49) coincides with the definition of two factors interaction used in the statistical field of Design of Experiments (Wu, 2015). If this term is positive than the interaction between i and j is profitable (i.e., synergistic). If it is negative then the interaction is disadvantageous (i.e., antagonistic). Thus, the intuition of Eq. (49) is that one averages this interaction index for all possible coalitions to which the subgroup s belongs. Note also that, when $s = \{i\}$, the Shapley-Owen value (48) reduces to the Shapley value (1). Recently, Lundberg et al. (2020) adopt the two-features Shapley-Owen value in (49) for the explanation of tree-based machine learning models.

Analogously to the Shapley value in Eq. (3), the Shapley-Owen effect can be also expressed in terms of the Moebius inverse by interpreting the group s as a single player (the team acts as one player).

Theorem 13 (*Grabisch & Roubens, 1999*)**.** *Given the value function v, the Shapley-Owen value of the feature group s can be written as*

$$Sh_s^v = \sum_{u \supseteq s} \frac{1}{|u|-|s|+1} m(u), \tag{50}$$

where m is the Moebius transform of v.

[2] Note that we are developing a parallel global framework to the local framework of Section 4.

The Shapley-Owen value is based on the following game-theoretic axioms, that differ from the ones at the basis of the Shapley value (Grabisch & Roubens, 1999):

1. (Linearity) $Sh_s^{v+w} = Sh_s^v + Sh_s^w$ for every $s \subseteq N$ and for any value function v and w.
2. (Dummy) If i is a dummy feature for v, then $Sh_i^v = v(\{i\})$ and $Sh_{s\cup\{i\}}^v = 0$ for every $s \subseteq N \setminus \{i\}$ with $s \neq \emptyset$.
3. (Symmetry) For all v and for all permutations π on N, $Sh_s^v = Sh_{\pi s}^{\pi v}$, where the game πv is defined by $\pi v(\pi s) = v(s)$, where $\pi s = \{\pi(i), i \in s\}$ for all $s \subseteq N$.
4. (Recursivity) ϕ^v obeys the following recurrence formula for every $s \subseteq N, |s| > 1$, any v and any $j \in s$: $Sh_s^v = Sh_{s\setminus\{j\}}^{\rho_j} - Sh_{s\setminus\{j\}}^{v^{N\setminus\{j\}}}$ where $\rho_j(s) = v(s \cup \{j\}) - v(\{j\})$ and $v^{N\setminus\{j\}}$ denotes the value of the game on $N \setminus \{j\}$ features.

Similarly to the procedure adopted for the Shapley value, in the next subsections we specify Shapley-Owen values for alternative value functions.

7.2. Finite-change Shapley-Owen index

In this section, we consider the definition of finite-change Shapley-Owen values for the finite-change value function $v(z) = \underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$. We recall that this value function quantifies the impact of the ith feature in the finite change Δt caused by the features shifting from $\mathbf{x}^0$ to $\mathbf{x}$. Let $Sh_s^{\mathbf{x}^0\to\mathbf{x}}$ denote the finite-change Shapley-Owen value of group s. We can characterize the finite-change Shapley-Owen values as follows.

Theorem 14. *Assume that $v(z) = \underline{\tau}_z^{\mathbf{x}^0\to\mathbf{x}}$. Then,*

$$Sh_s^{\mathbf{x}^0\to\mathbf{x}} = \sum_{u\supseteq s} \frac{\phi_u^{\mathbf{x}^0\to\mathbf{x}}}{|u| - |s| + 1}, \tag{51}$$

where ϕ_u is a finite-difference effect in (13).

Corollary 15. *If f is Δ-monotone, then*

$$Sh_s^{\mathbf{x}^0\to\mathbf{x}} \geq 0$$

for all $s \subseteq \{1, 2, \dots, n\}$.

We next provide a characterization of $Sh_s^{\mathbf{x}^0\to\mathbf{x}}$ at the infinitesimal scale.

Theorem 16. *Consider the finite-change Shapley-Owen value of the subgroup s. Then,*

$$Sh_s^{\mathbf{x}^0\to\mathbf{x}} \approx \frac{\partial^{|s|} f(\mathbf{x}^0)}{\partial x_{i_1} \cdots \partial x_{i_s}} \Delta x_{i_1} \cdots \Delta x_{i_s} \tag{52}$$

for $\mathbf{x} \to \mathbf{x}^0$ and for every $s \subseteq \{1, 2, \dots, n\}$.

The approximation in (52) is consistent with the notion of Shapley-Taylor interaction index defined in Dhamdhere et al. (2020). Precisely, Dhamdhere et al. (2020) consider the lth order Taylor series approximation of the model $f(\mathbf{x})$ and prove that the Shapley-Taylor interaction index for feature group s, ShT_s, with $|s| < l$, is the s-order partial derivative

$$ShT_s = \frac{\partial^{|s|} f(\mathbf{x}^0)}{\partial x_{i_1} \cdots \partial x_{i_s}}. \tag{53}$$

Note that if the input–output mapping is Δ-monotone, by (D.1) the partial derivatives (and thus the $Sh_s^{\mathbf{x}^0\to\mathbf{x}}$ for small changes) are positive: the positivity of the mixed partial derivatives is the assumption on the target function f in Dugas et al. (2009).

At the infinitesimal scale, $Sh_s^{\mathbf{x}^0\to\mathbf{x}}$ in (52) can be also connected to the crossed derivative-based global sensitivity measure of Roustant et al. (2014). This latter sensitivity measure is defined as (Roustant et al., 2014)

$$\zeta_s = \mathbb{E}\left[\left(\frac{\partial^{|s|} f(\mathbf{X})}{\partial X_{i_1} \cdots \partial X_{i_s}}\right)^2\right]. \tag{54}$$

The measure (54) extends to groups the derivative-based measure in (32) of Sobol' and Kucherenko (2009). Thus, rewriting Eq. (52) we find

$$\zeta_s \approx \mathbb{E}\left[\left(\frac{Sh_s^{\mathbf{X}\to\mathbf{X}+\Delta\mathbf{X}}}{\Delta X_{i_1} \cdots \Delta X_{i_s}}\right)^2\right] \tag{55}$$

for small values of $\Delta\mathbf{X}$. This equation is the extension of (36) to the case of feature groups.

7.3. Variance-based Shapley-Owen effects

We now consider the Shapley-Owen values (48) when the value function is the Sobol' index $v(u) = \underline{\tau}_u^2$. Rabitti and Borgonovo (2019) call the resulting interaction index the Shapley-Owen effect, in analogy with the terminology for Shapley effect (Song et al., 2016). The Shapley-Owen effects are a promising tool for quantifying interaction effects in the presence of feature dependence. These indices can be interpreted in terms of the explanatory power of the features when taken together as a group. Plischke et al. (2021) propose an algorithm for the computation of Shapley-Owen effects.

When features are independent, we can represent Shapley-Owen effects as Rabitti and Borgonovo (2019)

$$Sh_s^{VB} = \sum_{u\supseteq s} \frac{\sigma_u^2}{|u| - |s| + 1}. \tag{56}$$

This leads to the bracketing inequality for Shapley-Owen effects:

$$\sigma_s^2 \leq Sh_s^{VB} \leq \Upsilon_s^2, \tag{57}$$

where $\Upsilon_s^2 = \sum_{u\supseteq s} \sigma_u^2$ is the superset importance measure of Liu and Owen (2006) and it quantifies the global impact of the feature subgroup s in all higher-order Sobol' interaction terms. We now show a further characterization of Shapley-Owen effects in terms of finite changes (13):

Theorem 17. *Assume that f is L^2–integrable and that features are independent. Then, we have*

$$Sh_s^{VB} = \mathbb{V}\left[\mathbb{E}\left[\sum_{u\supseteq s} \frac{\phi_u^{\mathbf{X}^0\to\mathbf{X}}}{\sqrt{|u| - |s| + 1}}\right]\right]. \tag{58}$$

Eq. (58) has the same operational implication as (45). With the finite-change terms (13) evaluated at multiple points, one can construct both finite-change Shapley (17) and Shapley-Owen (51) values. From the same finite-change effects, by Theorems 11 and 17 the analysts can obtain the Shapley and Shapley-Owen variance-based effects and derivative-based global sensitivity measures (Eqs. (36) and (55)). We can generalize Theorem 12 of Mase et al. (2020) to Shapley-Owen effects.

Theorem 18. *Assume that f is L^2–integrable. Then, it holds*

$$Sh_s^{VB} = \mathbb{E}\left[Sh_s^{SC}(\mathbf{X})\right], \tag{59}$$

where the expectation is taken with respect to $\mathbf{X}$.

This new characterization is useful for the computation of Shapley-Owen effects. It shows that the given-data estimation procedure for the Shapley effects proposed by Mase et al. (2020) can be extended to interaction quantification. The accuracy and robustness of the numerical estimation of these indices constitute an open area of future research.

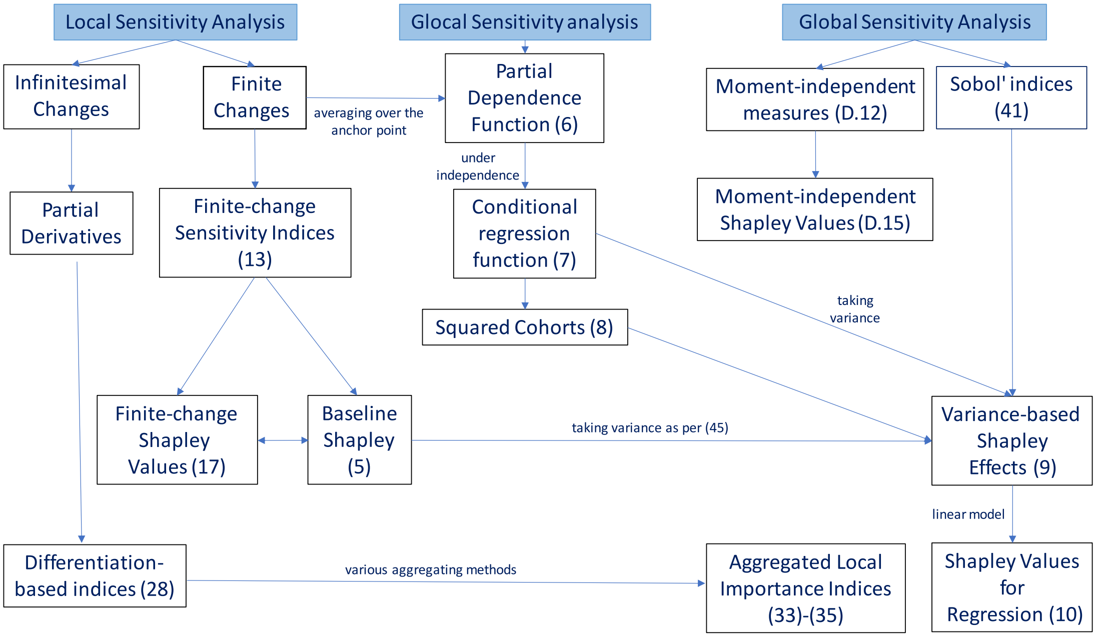

Local Sensitivity Analysis
Glocal Sensitivity analysis
Global Sensitivity Analysis
Infinitesimal Changes
Finite Changes
averaging over the anchor point
Partial Dependence Function (6)
Moment-independent measures (D.12)
Sobol' indices (41)
Partial Derivatives
Finite-change Sensitivity Indices (13)
under independence
Conditional regression function (7)
Moment-independent Shapley Values (D.15)
Squared Cohorts (8)
taking variance
Finite-change Shapley Values (17)
Baseline Shapley (5)
taking variance as per (45)
Variance-based Shapley Effects (9)
linear model
Differentiation-based indices (28)
various aggregating methods
Aggregated Local Importance Indices (33)-(35)
Shapley Values for Regression (10)

Fig. 1. Graphical summary of the hierarchy of value functions and corresponding Shapley (Shapley-Owen) values. The double arrow between the Baseline Shapley and Finite-Change Shapley signifies their equivalence as per Theorem 4.

7.4. An overview and the glocal scale

Fig. 1 provides a graphical illustration of the proposed hierarchy of value functions for Shapley and Shapley-Owen values. The left part displays local methods, the middle part glocal methods, and the right part global methods. Arrows denote their connections and numbers refer to the corresponding equation in the text. At a local level the analysis can be carried out on an infinitesimal or finite scale. In the former, we consider small perturbations of the features around their base case value. The sensitivity indices are gradients or differential sensitivities [Eq. (28)]. In the latter, we consider variations of the model across two entries. These can be seen as scenarios in sensitivity analysis. The sensitivity measures are finite-change indices in Eq. (13). Finite-change indices emerge from the exact decomposition of the change in model prediction across the two scenarios in 2^n terms. They are related to baseline Shapley values in Eq. (17) by Theorem 4 (the equivalence is signaled by the double arrow).

Finite change indices are also related to partial dependence functions, that are at the basis of the value function in Eq. (6) and, under independence, are equivalent to the conditional regression value function in Eq. (7). Squaring this value function one obtains the value function in Eq. (8) at the basis of cohort Shapley effects $Sh_i^{\mathrm{SC}}(\mathbf{x})$ in Eq. (47), whose value depends on the anchor point $\mathbf{x}$. Note that averaging these squared cohort Shapley values we obtain variance-based Shapley values in Eq. (9), which are global indices.

However, displaying how the Shapley values vary as a function of the reference individual provides additional insights into the model behavior, and can be used to create regional representations of sensitivity information. This would otherwise remain hidden if the local Shapley values are simply combined to obtain global explanations. This sensitivity analysis, which is intermediate between a local and a global scale, is called glocal.

The right part of the figure refers to the global level. In a global sensitivity analysis, one inspects the model at several locations in the feature space. Global sensitivity measures, either variance-based (Sobol' indices) [Eq. (40)] or moment-independent [Eq. (D.12)] can be both associated with Shapley values. In particular, variance-based Shapley effects [Eqs. (9) and (43)] have been discussed in Owen (2014), Owen and Prieur (2017), while moment-independent Shapley values are discussed in Sarazin et al. (2020).

8. Application

In this section, we explore the integration of local and global insights using Shapley values and quantify interactions at different scales using Shapley-Owen values analyzing the dataset Medical Insurance Premium Predictions available at https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction. In discussing the managerial side of our results, we consider the perspective of three stakeholders: an insurance company, a regulator and the individual.

This dataset contains 986 individual medical premiums based on 10 features related to the policyholder's characteristics: age (continuous), diabetes (discrete), blood pressure problems (discrete), any organ transplants (discrete), any chronic diseases (discrete), height (continuous), weight (continuous), any known allergies (discrete), history of cancer in family (discrete), number of major surgeries (discrete). We have fitted a feed-forward neural network with 2 layers with 10 and 1 nodes respectively using the Matlab2019 Neural Network package. We split the data into 75% for training and 25% for testing. The R^2 coefficient is equal to 0.93 on the training data and 0.80 on the testing data. From now on we consider this ML model.

To investigate which features are responsible for moving the premiums from the minimum to the maximum at any policyholders' age we use the local Shapley values in Eq. (17). Specifically, for any age we consider the minimum and maximum premium as base case ($\mathbf{x}^0$) and sensitivity case ($\mathbf{x}^1$), respectively. Namely, at a given age we set as base case the feature values of the policyholder paying the lowest premium, and as sensitivity case those of the policyholder paying the highest premium. We then compute the finite-change Shapley values via Eq. (17) shifting the features across the two individuals, and we repeat it for all ages. Note that, since two policyholders with the same age are always considered for the finite changes, age does not appear as a model input in this analysis. Results are shown in Fig. 2.

In Fig. 2, we report on the horizontal axis the policyholder's age, and on the vertical axis the magnitude of Shapley values. At each age, nine dots are plotted, corresponding to the finite-change Shapley values that distribute the difference between the maximum and minimum

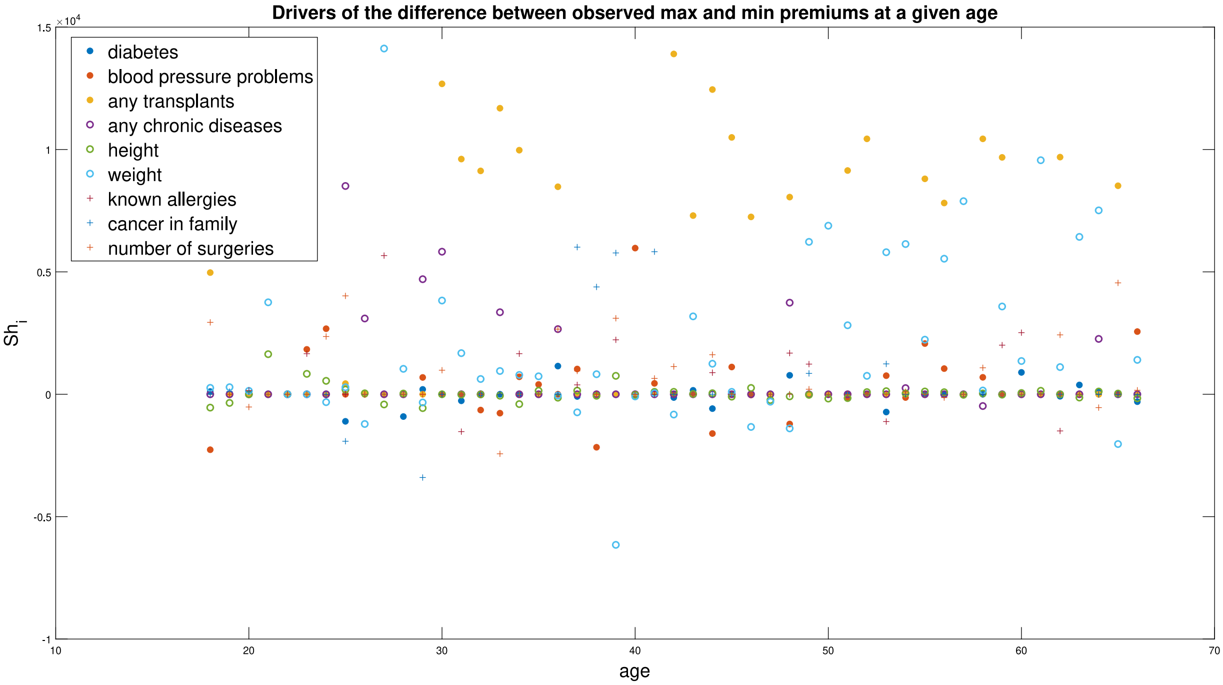


Fig. 2. Finite-change Shapley values for the difference between the maximum and the minimum of the observed premiums as a function of policyholders' age (in years).

premiums among the nine features of the policyholders. The sum of these Shapley values is positive at any age, as it reflects the difference in premiums. Fig. 2 shows that the presence of transplants (yellow asterisk), chronic diseases (purple circle) and the body weight (blue circle) are generally the most important drivers of the difference between the maximum and minimum premiums for any age. Notably, Fig. 2 evidences a clear pattern in the impact of these features across different ages. In particular, the presence of transplants is the most important feature in driving the premium up, but only when age is greater than 30. Weight is the second most important feature in increasing the premiums if age is greater than 50. For lower ages, the presence of a chronic disease becomes an important feature. Transplants and chronic diseases are conditions that require ongoing medical care and treatment, which can result in higher healthcare costs and thus a higher premium. Similarly, body weight is a known risk factor for a variety of health problems, and it is well documented that individuals who are overweight or obese show an increased likelihood of developing health problems. Thus, these findings are consistent with medical knowledge.

Regarding managerial insights, individual finite change Shapley values can be used to better inform the single policyholder, who gains insights into how their intrinsic features contribute to the final algorithmic decision. First, the policyholder can appreciate whether the decision of the algorithm was based on fair or unfair attributes. Second, they can determine whether the important features are actionable. To illustrate, if weight appears to be a relevant factor for a policyholder of age greater than 50, the person can decide whether or not to take actions to reduce their body weight, towards decreasing the insurance premium. Also, these results can be used by the insurance company and the regulator: in analyzing the case of a particular policyholder, they need insights to understand and explain the rationale of that specific algorithmic decision. Moreover, the use of finite change Shapley values can support the insurance company that can develop tailored health insurance policies specific for the needs of different customer segments. For example, they could provide optional coverage for certain pre-existing conditions or lifestyle-related benefits, allowing customers to personalize their plans.

Moving towards a glocal scale, we now consider the explanations of all policyholders against the average individual. As base case, we set $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{10})$, which represents the average policyholder with the mean feature values of all policyholders. As Mase et al. (2020) highlight, this average policyholder is not necessarily a real individual in the data. We then repeat the previous analysis calculating the finite-change Shapley values with respect to this new base case. Fig. 3 displays the results. Fig. 3 shows that age (blue circles) is the most important feature in affecting premiums for policyholders younger than 30. The effect of age significantly reduces insurance premiums for this age group. However, for policyholders over 50, the effect of age increases the insurance premium. This is intuitive, as younger people are less likely to require healthcare and therefore incur lower expenditures compared to older individuals. Additionally, the presence of any transplants (purple circles) significantly increases the individual premium. In contrast, for policyholders between 25 and 49, the presence of chronic diseases (green circles) and a family history of cancer (red asterisk) increase premiums. Moreover, Fig. 3 displays a clear pattern of how individual determinants change across different ages. At all ages, the most influential feature driving the increase in the premium is having any transplant. For policyholders aged between 18 and 24, family cancer history has the next significant impact, while the presence of any chronic disease becomes prominent for those aged 25 to 33. These two features then alternate between the second and third positions until the age of 49, after which policyholder's age is the second most important driver of the premium increase. Knowledge of how a specific feature impacts premiums at different ages can inform managerial decisions, such as designing and pricing individual health-insurance contracts.

As a subsequent glocal sensitivity analysis, we compute the individual squared cohort Shapley values in Eq. (8). We estimate them in two ways: by using a ML model (the neural network model previously fitted) and directly through the cohort method of Mase et al. (2020) - see Appendix 3 for a description. Results are plotted in Fig. 4. Fig. 4 consists of two panels: the lower panel shows the squared-cohort Shapley values estimated using the neural network model, while the upper panel displays the cohort-based estimates. Both panels evidence a similar pattern, indicating that the ML model does not distort the indications coming from the original data. The blue circles in both panels highlight the impact of age, showing that for policyholders younger than 30, all blue circles are significantly above all other Shapley values.

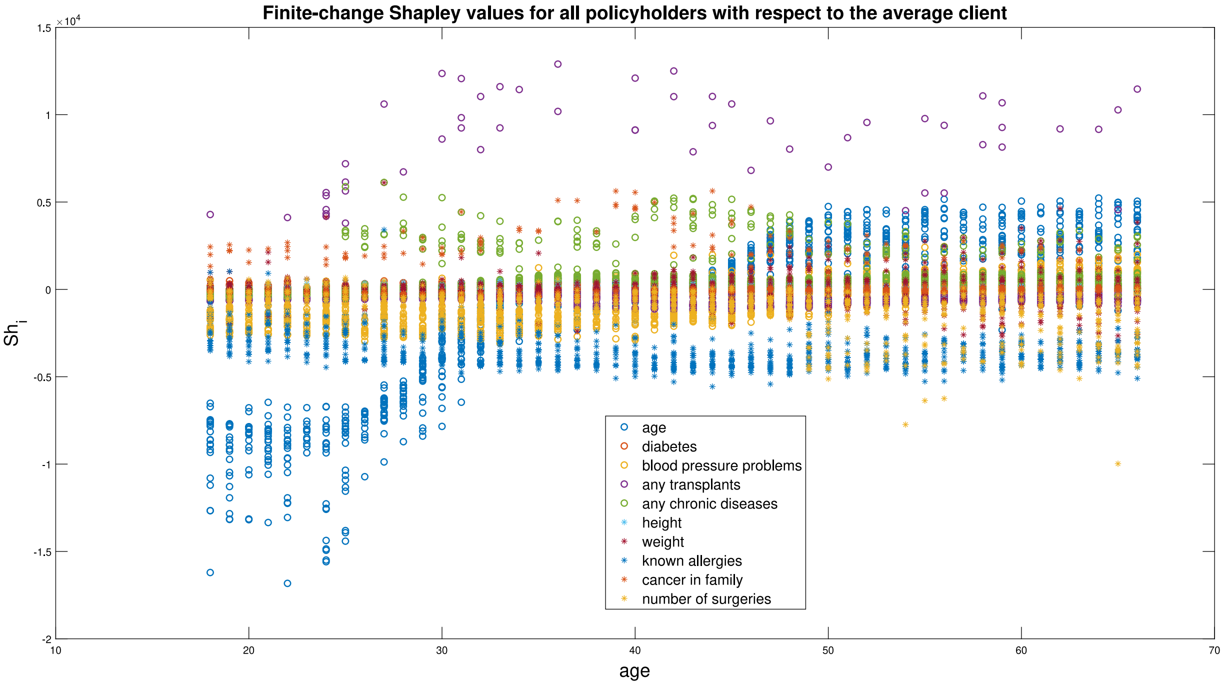


Fig. 3. Finite-change Shapley values for individual policyholders as function of their age, computed with the neural network. The base point is the average policyholder.

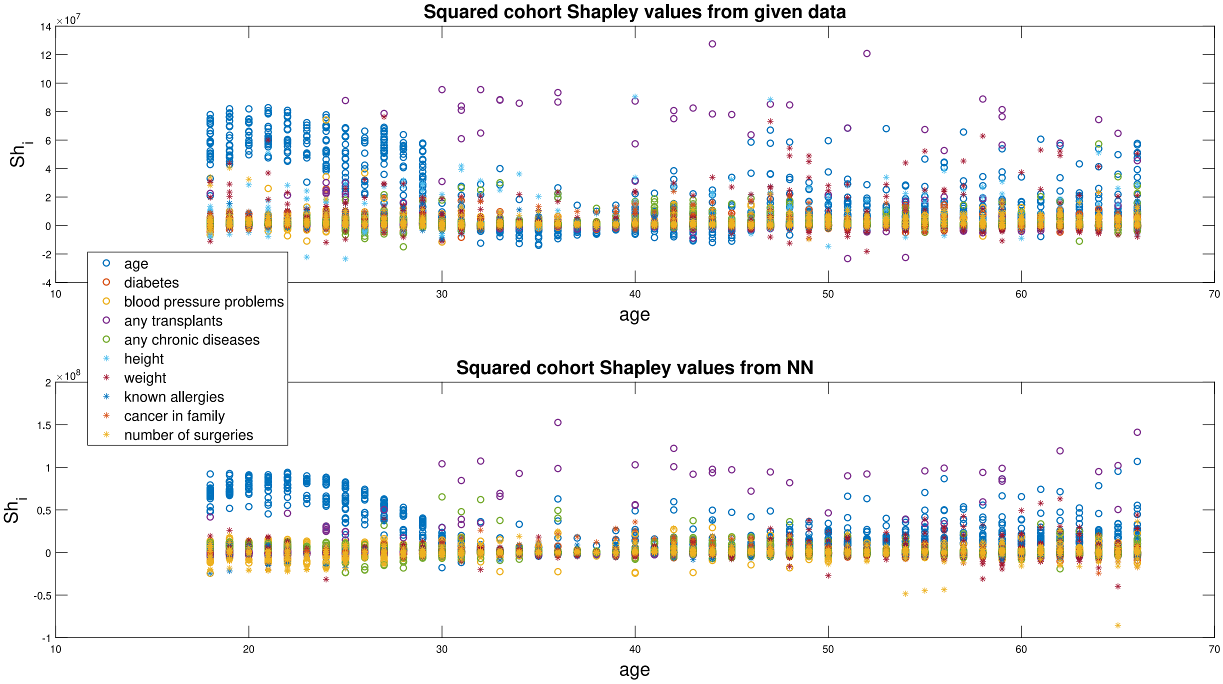
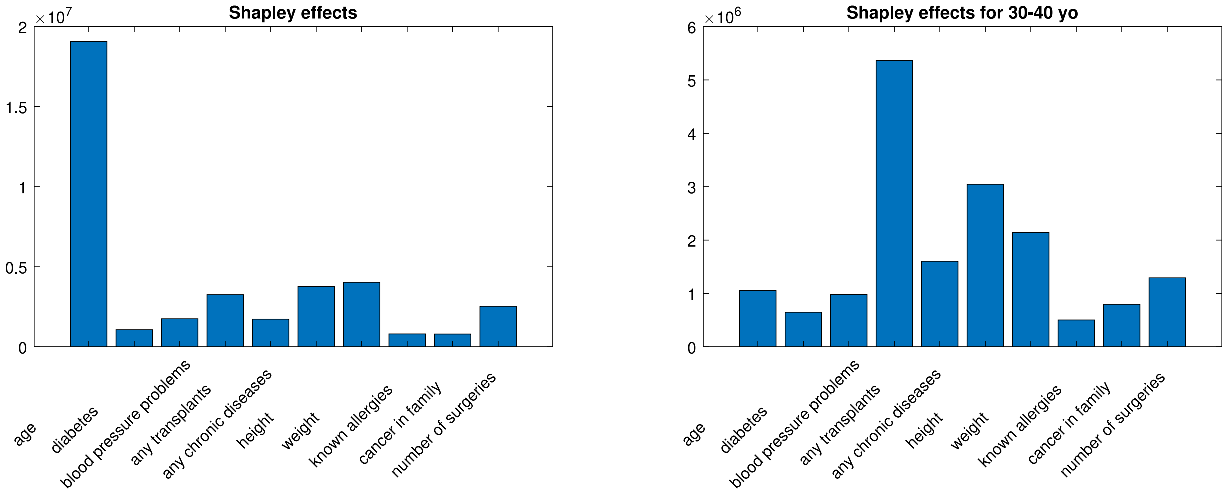


Fig. 4. Squared-cohort Shapley values for individual policyholders as function of their age, estimated from cohorts (upper panel) or with neural network (lower panel).

For ages between 30 and 40, the presence of any transplants and any chronic diseases have the most significant impact on the insurance premium. We can also compare the insights obtained from the two glocal approaches. Figs. 3 and 4 are consistent, although Fig. 3 provides additional information on the direction of the effects, indicating whether a feature increases or decreases the insurance premium.

Regarding managerial insights, results at the glocal level are now leading insights at the population level. They are valuable for insurance supervisors looking at their portfolio. Having access to the primary drivers of the predictive model enables regulators to scrutinize for potential discriminatory effects on a large scale, evidence, for instance, by a significant importance of variables like “race” or “gender”, — however, these variables are not present in this dataset. The glocal sensitivity analysis, which provides insights regarding individual policyholders compared to similar insured individuals in the portfolio, is highly relevant for the insurance company. For example, if the company discovers that a group of policyholders with certain characteristics poses lower risk compared to others, it might offer them more competitive rates or additional benefits to incentivize their loyalty. Similarly, identifying individual policyholders who present significantly different



(a) Shapley effects.

(b) Shapley effect for the age group 30-40 yo.

Fig. 5. Shapley effects estimated with cohorts for the whole data (left panel) and for the age group 30–40 yo (right panel).

risk levels compared to their counterparts in the portfolio allows the company to implement targeted risk management actions.

We then investigate the global drivers of the premiums. According to Theorem 12, the average of the squared-cohort Shapley values produces the Shapley effects Sh^{VB}, which are displayed in Fig. 5(a). This figure shows that age is globally the key-driver of medical premiums. With reference to Theorem 12, we can gain additional insights. The globally most important feature might not necessarily be the most relevant for a specific subgroup of individuals. Therefore, instead of averaging the squared-cohort Shapley values over all policyholders, we consider subjects in specific age groups, such as those between 30 and 40 years old, and compute the Shapley effects for this subgroup. Results are shown in Fig. 5(b). This plot shows that for this specific age group the most important feature is the presence of transplants.

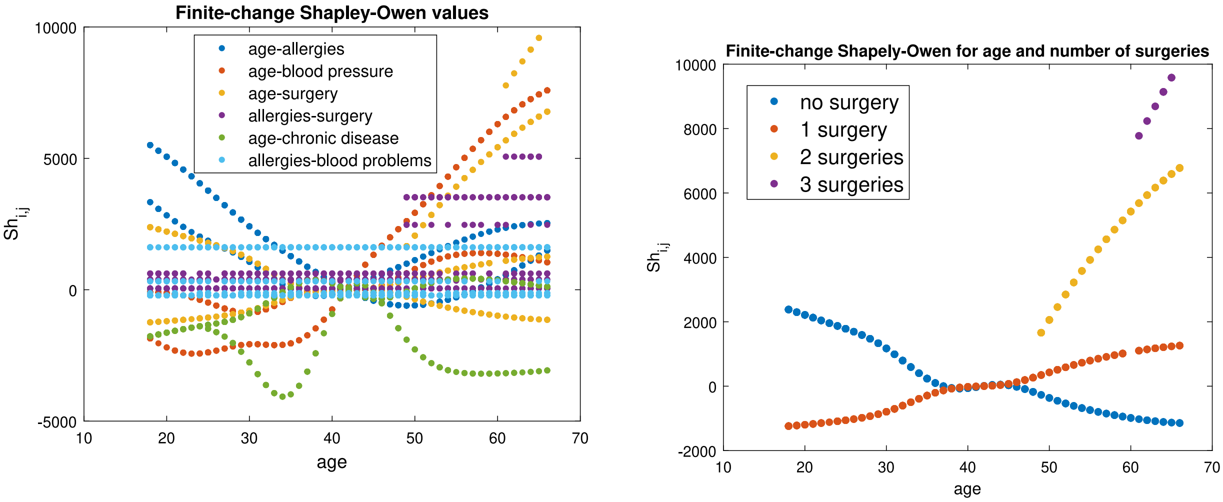
These indications promote transparency and assist modelers in constructing and refining pricing models by monitoring their primary drivers. For the modeler, they are appropriate to answer the question of what are the dataset-level drivers of the algorithmic response and are especially useful in the initial modeling phases, when analysts can compare these dataset indications on model predictions with results obtained from other measures of importance calculated at the dataset level. For supervisors or insurance companies, they can be used to compare portfolios. An insurer may compare two different portfolios which they regarded as similar to have a confirmation of this intuition or being open to discover that different features are instead relevant for the portfolios. Detailed understanding of variable contributions to insurance premiums enables more accurate risk management because the company becomes aware of the risk drivers and can better prioritize actions. For example, the company might decide to obtain a larger rating class without changing significantly the individual premiums by ignoring subgroups generated by the least important pricing variable. Larger classes are more stable, as there is a greater solidarity effect among the policyholders.

We now come to the interaction analysis conducted with Shapley-Owen effects. Understanding how features interact is essential for understanding how the insurance pricing model works.

We start considering the finite-change Shapley-Owen values with the average policyholder $\bar{\mathbf{x}}$ as base case. We plot the most relevant pairwise finite-change Shapley-Owen values in Fig. 6(a) as a function of age. Fig. 6(a) indicates that age and any blood problems have a joint negative effect on the premium for younger ages but a positive effect for older ages. Moreover, the joint effect of age and number of surgeries greatly increases the premium for older ages. As shown later in Fig. 7(a), this latter joint effect is globally the most relevant. Specifically, Fig. 6(b) shows that being older than 50 and having had at least 2 surgeries is the pair of features with the largest local effect on the difference in insurance premium with respect to the average policyholder.

Moving to the global scale, we consider the variance-based Shapley-Owen effects which can be interpreted in terms of explanatory power of two (or more) features taken together. Fig. 7(a) displays the Shapley-Owen effects between age and all the other features computed via Eq. (59). Fig. 7(a) shows that age and number of surgeries have a very negative explanatory power when considered together, meaning that these features provide globally redundant information to the pricing model. By Theorem 18, the Shapley-Owen effects can be estimated as the average of the squared cohort Shapley-Owen values. Thus, to investigate the Shapley-Owen effects between age and number of surgeries, we plot the squared cohort Shapley-Owen values between these two features in Fig. 7(b). This figure shows that the joint contributions of age and two or three surgeries are highly negative at ages greater than 60. Thus, for these ages, age and the occurrence of two or three surgeries provide redundant information for insurance pricing. Figs. 6(b) and 7(b) allow us to highlight the similarities and differences between finite-change and squared cohort Shapley-Owen values. Finite-change Shapley-Owen values require the use of an ML model, such as a neural network in this study, and can be interpreted to understand the financial impact of local joint effects on the difference in premiums between the base case and the sensitivity case. On the other hand, the cohort approach can be used to compute Shapley-Owen indices from given data, and these indices can help determine whether two features provide redundant or enhanced information. Figs. 6(b) and 7(b) show that having undergone two or three surgeries and being over 50 years of age lead to an increase in premium. Additionally, these two features provide globally similar information regarding the variability of the premiums.

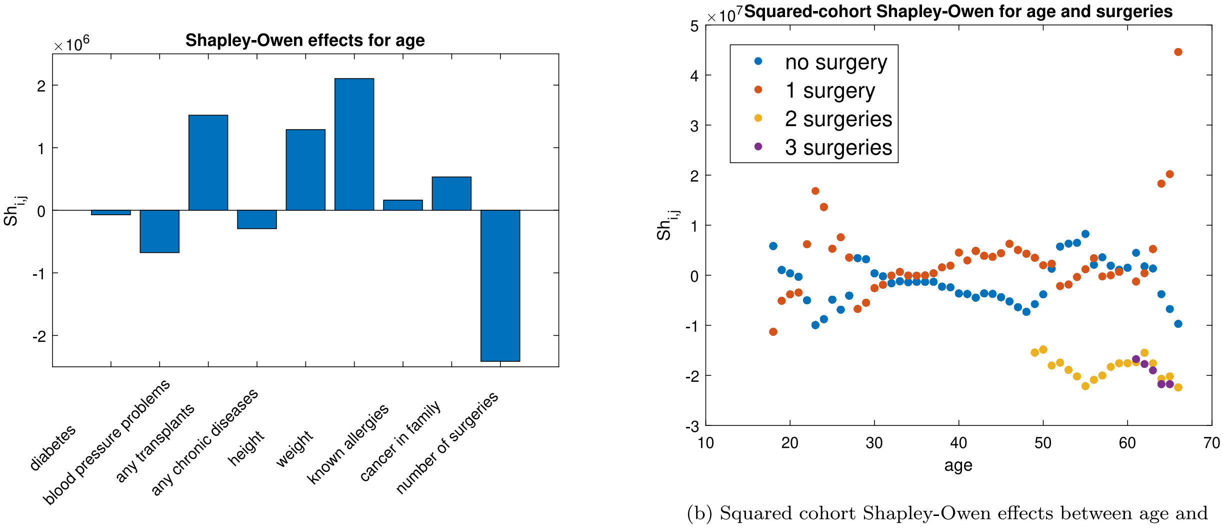
An interaction analysis can assist the insurance company as follows. Identifying an important joint contribution that significantly increases the premium can lead the insurance company to create tailored contracts. Moreover, understanding how risk factors jointly influence insurance premia allows the company to take targeted preventive measures to mitigate the effects associated with certain feature combinations.



(a) Finite-change Shapley-Owen values for individual policyholders with $\bar{\mathbf{x}}$ as base case value, computed with the neural network.

(b) Focus on finite-change Shapley-Owen values between age and number of surgeries.

Fig. 6. Finite change Shapley-Owen values. Some values are plotted in the left panel as function of age, while the right panel depicts a focus on those values between age and number of surgeries.



(a) Shapley-Owen effects of age.

(b) Squared cohort Shapley-Owen effects between age and any number of surgeries.

Fig. 7. Interaction analysis for the age, which is the feature with highest interaction effects. Shapley-Owen effects for are displayed on the left panel, while squared Shapley-Owen values for the interaction between age and number of surgeries is displayed on the right. Both are estimated using the cohort method applied to the predicted premiums with the neural network.

9. Conclusions

The paper has analyzed the alternative formulations of Shapley explanations developed in the literature. A take-away from our analysis is that there is no universal strategy to analyze an ML model. We then present in Fig. 8 a flowchart that might be useful towards a systematic selection.

Given an available dataset, after fitting a machine learning model, the first choice is establishing the goal of the analysis. If the analyst is interested in explaining the difference between two individual predictions (first decision node), then the we are at a local scale and baseline Shapley that descend from the value function in (5) are the appropriate sensitivity indices. Otherwise, if the analysts' goal is the overall importance of the features, then a dataset-level formulation becomes appropriate. We are at a global level and Shapley explanations such as those in Eqs. (8) or (9) are appropriate. If none of these is the goal, a bridge is the repeated calculation of local importance values at multiple locations in the dataset. For instance, the SHAPs subroutine yields as many Shapley values for a given feature as many are the points in the dataset. Note that Senoner et al. (2022) take the average the absolute values of the Shapley effects to obtain an overall dataset importance. However, there is no unique way of aggregating the repeated evaluation of sensitivities: one could also use the largest magnitude or the upper 95 quantile. This way of proceeding mixes a local formulation aimed at individual explanations with a global explanation and may result in an ad-hoc aggregation. While this is certainly possible, we would still recommend comparing the results of such aggregation procedure with indications from dataset-level indices, either Shapley values or measures of statistical association (see Borgonovo et al., 2023).

Moreover, aggregating results at the local scale might undermine the derivation of insights. A compromise between the local and global approaches is attained in a setting that we have called *glocal*. By visualizing glocal information, we gain precious additional regional

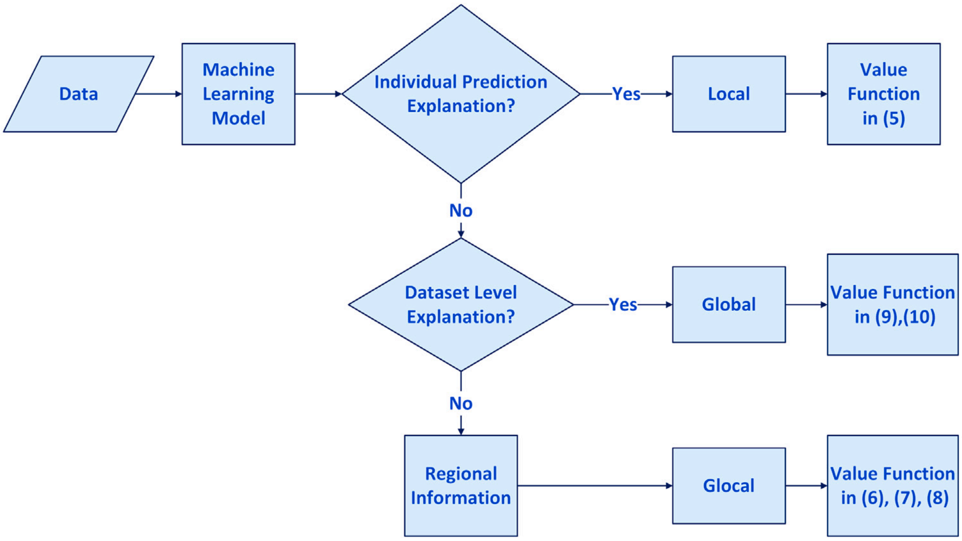

Data
Machine Learning Model
Individual Prediction Explanation?
Yes
Local
Value Function in (5)
No
Dataset Level Explanation?
Yes
Global
Value Function in (9),(10)
No
Regional Information
Glocal
Value Function in (6), (7), (8)

Fig. 8. A flowchart guiding the choice of alternative Shapley Value formulations.

insights. Insights on how the model responds for specific subgroups of features or realizations augment its explainability and deepen the analyst's understanding, as drivers of the model behavior might not be the same for specific subgroups of subjects or specific feature variations.

Our investigation shows that the insights yielded by investigations at the different scales are complementary rather than overlapping and may be useful to different stakeholders. Individual policyholders focus naturally on single predictions, while an insurance company may be interested in managing risk at a portfolio level.

Similar considerations are valid for Shapley-Owen indices that quantify the joint feature contributions to the model predictions and that can be also computed at any scale. Their calculation yields indications on interactions that can be used to improve risk management. In our application we have seen that determining whether the simultaneous variation in two or more individuals' characteristics amplifies or dampens their individual effects helps in formulating tailored insurance contracts.

While the Shapley values can provide many insights into the explanation of decision models, as we have shown, their main limitation lies in their computational cost. Indeed, in the absence of closed form expressions, the value functions of every possible coalition need to be estimated, resulting in an exponential cost. Therefore, research into algorithmic and statistical methods for calculating and approximating Shapley values becomes fundamental to estimate them in high-dimensional contexts. Also, at the dataset level, alternative and less computationally heavy methods may produce robust indications about feature importance. One can recall, for instance, measures of statistical association based on alternative rationales, among which the new correlation coefficient of Chatterjee (2021) or the Wasserstein dependence measures of Nies et al. (2023) and Wiesel (2022). A comparison of these methods and Shapley values is then an open future research avenue.

CRediT authorship contribution statement

Emanuele Borgonovo: Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Elmar Plischke:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Giovanni Rabitti:** Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to thank the editor, Professor Roman Slowinski, and the three anonymous reviewers for several perceptive and constructive observations. Part of this work has been written while Giovanni Rabitti was visiting the Bocconi Institute for Data Science and Analytics through the "UK-Italy Trustworthy AI Visiting Researcher Programme" of the Alan Turing Institute.

Appendix A. Supplementary data

All proofs are contained in the online supplementary material. The codes for reproducing the simulations can be found at https://github.com/giovanni-rabitti/many-shapley.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.06.023.

References

Aas, K., Jullum, M., & Løland, A. (2019). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. arXiv e-prints arXiv:1903.10464.

Ahmed, A., Topuz, K., Moqbel, M., & Abdulrashid, I. (2024). What makes accidents severe! explainable analytics framework with parameter optimization. *European Journal of Operational Research*, *317*, 425–436.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to Explain Individual Classification Decisions. *Journal of Machine Learning Research*, *11*, 1803–1831.

Balog, D., Bátyi, T. L., Csóka, P., & Pintér, M. (2017). Properties and comparison of risk capital allocation methods. *European Journal of Operational Research*, *259*(2), 614–625.

Baucells, M., & Borgonovo, E. (2013). Invariant Probabilistic Sensitivity Analysis. *Management Science*, *59*(11), 2536–2549.

Becker, W. E., Tarantola, S., & Deman, G. (2018). Sensitivity analysis approaches to high-dimensional screening problems at low sample size. *Journal of Statistical Computation and Simulation*, *88*(11), 2089–2110.

Benoumechiara, N., & Elie-Dit-Cosaque, K. (2019). Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms. In *ESAIM: ProcS, vol. 65* (pp. 266–293).

Bergantiños, G., Groba, C., & Sartal, A. (2023). Applying the Shapley value to the tuna fishery. *European Journal of Operational Research*, *309*(1), 306–318.

Borgonovo, E. (2010). Sensitivity analysis with finite changes: An application to modified EOQ models. *European Journal of Operational Research*, *200*(1), 127–138.

Borgonovo, E., & Apostolakis, G. E. (2001). A new importance measure for risk-informed decision making. *Reliability Engineering & System Safety*, *72*(2), 193–212.

Borgonovo, E., Ghidini, V., Hahn, R., & Plischke, E. (2023). Explaining classifiers with measures of statistical association. *182*, (pp. 1–20).

Borgonovo, E., & Rabitti, G. (2023). Screening: From tornado diagrams to effective dimensions. *European Journal of Operational Research*, *304*, 1200–1211, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130919569&doi=10.1016%2Fj.ejor.2022.05.003&partnerID=40&md5=4316dba3b71c9c2a6164283e957bcb8a.

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, *22*(10), 1509–1518.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, *116*(536), 2009–2022.

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*, 357–372.

Cheng, X., Li, G., Skulstad, R., Chen, S., Hildre, H. P., & Zhang, H. (2019). A neural-network-based sensitivity analysis approach for data-driven modeling of ship motion. *IEEE Journal of Oceanic Engineering*, 1–11.

Cohen, S., Dror, G., & Ruppin, E. (2007). Feature selection via coalitional game theory. *Neural Computation*, *19*(7), 1939–1961.

Csóka, P., Illés, F., & Solymosi, T. (2022). On the Shapley value of liability games. *European Journal of Operational Research*, *300*(1), 378–386.

Daniels, H., & Velikova, M. (2010). Monotone and partially monotone neural networks. *IEEE Transactions on Neural Networks*, *21*(6), 906–917.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy* (pp. 598–617).

De Bock, K. W., Coussement, K., Caigny, A. D., Słowiński, R., Baesens, B., Boute, R. N., Choi, T. M., Delen, D., Kraus, M., Lessmann, S., Maldonado, S., Martens, D., Óskarsdóttir, M., Vairetti, C., Verbeke, W., & Weber, R. (2023). Explainable AI for Operational Research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, http://dx.doi.org/10.1016/j.ejor.2023.09.026, URL: https://www.sciencedirect.com/science/article/pii/S0377221723007294.

Dhamdhere, K., Agarwal, A., & Sundararajan, M. (2020). The Shapley Taylor Interaction Index. In *Proceedings of the 37th international conference on machine learning* (pp. 9259–9268).

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., & Garcia, R. (2009). Incorporating functional knowledge in neural networks. *Journal of Machine Learning Research*, *10*, 1239–1262.

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, *9*(3), 586–596.

Engelbrecht, A. P., Cloete, I., & Zurada, J. M. (1995). Determining the significance of input parameters using sensitivity analysis. In J. Mira, & F. Sandoval (Eds.), *From natural to artificial neural computation* (pp. 382–388). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fernández-Navarro, F., Carbonero-Ruz, M., Becerra Alonso, D., & Torres-Jiménez, M. (2017). Global Sensitivity Estimates for Neural Network Classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11), 2592–2604.

Florez-Lopez, R. (2007). Modelling of insurers' rating determinants. An application of machine learning techniques and statistical models. *European Journal of Operational Research*, *183*(3), 1488–1512.

Fock, E. (2014). Global Sensitivity Analysis Approach for Input Selection and System Identification Purposes: A New Framework for Feedforward Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(8), 1484–1495.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954.

Fu, R., Aseri, M., Singh, P. V., & Srinivasan, K. (2022). ''Un''Fair Machine Learning Algorithms. *Management Science*, *68*(6), 4173–4195.

Glasserman, P. (1990). *Gradient estimation via perturbation analysis*. Springer.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Grabisch, M., & Roubens, M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, *28*(4), 547–565.

Grömping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician*, *61*(2), 139–147.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(2), 137–152.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 93:1—-93:42.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, *19*(3), 293–325.

Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, *52*(1), 1–17.

Hong, L. J., & Liu, G. (2009). Simulating Sensitivities of Conditional Value at Risk. *Management Science*, *55*(2), 281–293.

Hong, L. J., & Liu, G. (2010). Pathwise Estimation of Probability Sensitivities Through Terminating or Steady-State Simulations. *Operations Research*, *58*(2), 357–370.

Hooker, G. (2004). Discovering Additive Structure in Black Box Functions. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 575–580). New York, NY, USA: ACM.

Horel, E., Mison, V., Xiong, T., Giesecke, K., & Mangu, L. (2018). Sensitivity based Neural Networks Explanations. arXiv e-prints arXiv:1812.01029.

Huettner, F., & Sunder, M. (2012). Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. *Electronic Journal of Statistics*, *6*, 1239–1250.

Iooss, B., & Prieur, C. (2019). Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol' indices, numerical estimation and applications. *International Journal of Uncertainty Quantification*, *9*(5), 493–514.

Janzing, D., Minorics, L., & Blöbaum, P. (2019). Feature relevance quantification in explainable AI: A causal problem. arXiv e-prints arXiv:1910.13413.

Kowalski, P. A., & Kusy, M. (2018a). Determining significance of input neurons for probabilistic neural network by sensitivity analysis procedure. *Computational Intelligence*, *34*(3), 895–916.

Kowalski, P. A., & Kusy, M. (2018b). Sensitivity Analysis for Probabilistic Neural Network Structure Reduction. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1919–1932.

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, *302*(1), 309–323.

Kuo, F. Y., Sloan, I. H., Wasilkowski, G. W., & Wozniakowski, H. (2010). On decompositions of multivariate functions. *Mathematics of Computation*, *79*, 953–966.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science*, *33*(1), 126–148.

Li, B., & Chen, C. (2018). First-Order Sensitivity Analysis for Hidden Neuron Selection in Layer-Wise Training of Networks. *Neural Processing Letters*, *48*(2), 1105–1121.

Lindelauf, R. H. A., Hamers, H. J. M., & Husslage, B. G. M. (2013). Cooperative game theoretic centrality analysis of terrorist networks: The cases of Jemaah Islamiyah and Al Qaeda. *European Journal of Operational Research*, *229*(1), 230–238.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*(4), 319–330.

Liu, R., & Owen, A. B. (2006). Estimating mean dimensionality of analysis of variance decompositions. *Journal of the American Statistical Association*, *101*(474), 712–721.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc..

Mase, M., Owen, A. B., & Seiler, B. B. (2020). Explaining black box decisions by Shapley cohort refinement. arXiv e-prints arXiv:1911.00467.

Molnar, C. (2018). *Interpretable machine learning*. Leanpub.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Naumzik, C., Feuerriegel, S., & Nielsen, A. M. (2023). Data-driven dynamic treatment planning for chronic diseases. *European Journal of Operational Research*, *305*(2), 853–867.

Nazemi, A., Baumann, F., & Fabozzi, F. J. (2022). Intertemporal defaulted bond recoveries prediction via machine learning. *European Journal of Operational Research*, *297*(3), 1162–1177.

Ni, J., Chen, B., Allinson, N. M., & Ye, X. (2020). A hybrid model for predicting human physical activity status from lifelogging data. *European Journal of Operational Research*, *281*(3), 532–542.

Nies, T. G., Staudt, T., & Munk, A. (2023). Transport dependency: Optimal transport based dependency measures. *v3*, (pp. 1–79).

Owen, G. (1972). Multilinear Extensions of Games. *Management Science*, *18*(52), 64–79.

Owen, A. B. (2014). Sobol' Indices and Shapley Value. *SIAM/ASA Journal on Uncertainty Quantification*, *2*(1), 245–251.

Owen, A. B., & Prieur, C. (2017). On Shapley Value for Measuring Importance of Dependent Inputs. *SIAM/ASA Journal on Uncertainty Quantification*, *5*(1), 986–1002.

Pesenti, S. M., Millossovich, P., & Tsanakas, A. (2021). Cascade Sensitivity Measures. *Risk Analysis*, *41*(12), 2392–2414.

Plischke, E., Rabitti, G., & Borgonovo, E. (2021). Computing Shapley Effects for Sensitivity Analysis. *SIAM/ASA Journal on Uncertainty Quantification*, *9*(4), 1411–1437.

Rabitti, G., & Borgonovo, E. (2019). A Shapley-Owen index for interaction quantification. *SIAM/ASA Journal on Uncertainty Quantification*, *7*(3), 1060–1075.

Rabitz, H., & Alis, Ö. F. (1999). General foundations of high - dimensional model representations. *Journal of Mathematical Chemistry*, *25*(2–3), 197–233.

Rota, G. C. (1964). On the foundations of combinatorial theory I. Theory of Möbius Functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *2*(4), 340–368.

Roustant, O., Fruth, J., Iooss, B., & Kuhnt, S. (2014). Crossed-derivative based sensitivity measures for interaction screening. *Mathematics and Computers in Simulation*, *105*, 105–118.

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.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*(none), 1–85.

Rüschendorf, L. (2013). *Mathematical risk analysis*. Berlin Heidelberg: Springer-Verlag.

Saltelli, A., Tarantola, S., & Campolongo, F. (2000). Sensitivity Analysis as an Ingredient of Modeling. *Statistical Science*, *15*(4), 377–395.

Sarazin, G., Derennes, P., & Morio, J. (2020). Estimation of high-order moment-independent importance measures for Shapley value analysis. *Applied Mathematical Modelling*, *88*, 396–417.

Scholbeck, C. A., Moosbauer, J., Casalicchio, G., Gupta, H., Bischl, B., & Heumann, C. (2023). Bridging the gap between machine learning and sensitivity analysis. *2312*, (pp. 1–14).

Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science*, *68*, 5704–5723.

Shapley, L. S. (1953). A Value for n-person Games. In H. W. Kuhn, & A. W. Tucker (Eds.), *Contributions to the theory of games* (pp. 307–317). Princeton University Press.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th international conference on machine learning* (pp. 1–9). Sydney: PMLR 70.

Sill, J. (1998). Monotonic Networks. In *Advances in neural information processing systems 10* (pp. 661–667).

Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, *1*(4), 407–414.

Sobol', I. M., & Kucherenko, S. (2009). Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, *79*(10), 3009–3017.

Sobrie, L., Verschelde, M., & Roets, B. (2023). Explainable real-time predictive analytics on employee workload in digital railway control rooms. *European Journal of Operational Research*.

Song, E., Nelson, B. L., & Staum, J. (2016). Shapley Effects for Global Sensitivity Analysis: Theory and Computation. *SIAM/ASA Journal on Uncertainty Quantification*, *4*(1), 1060–1083.

Štrumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications Using Game Theory. *Journal of Machine Learning Research*, *11*, 1–18.

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

Štrumbelj, E., Kononenko, I., & Robnik Šikonja, M. (2009). Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, *68*(10), 886–904.

Sundararajan, M., & Najmi, A. (2019). The many Shapley values for model explanation. arXiv e-prints arXiv:1908.08474.

Tsaih, R. H. (1999). Sensitivity analysis, neural networks, and the finance. In *IJCNN'99. international joint conference on neural networks. proceedings, vol. 6* (pp. 3830–3835).

Tsanakas, A., & Millossovich, P. (2016). Sensitivity Analysis Using Risk Measures. *Risk Analysis*, *36*(1), 30–48.

Wagner, H. M. (1995). Global Sensitivity Analysis. *Operations Research*, *43*(6), 948–969.

Wang, X. Z., Li, C. G., Yeung, D. S., Song, S., & Feng, H. (2008). A definition of partial derivative of random functions and its application to RBFNN sensitivity analysis. *Neurocomputing*, *71*(7), 1515–1526.

Wiesel, J. C. W. (2022). Measuring association with Wasserstein distances. *Bernoulli*, *28*(4), 2816–2832.

Wu, C. F. J. (2015). Post-Fisherian experimentation: from physical to virtual. *Journal of the American Statistical Association*, *110*(510), 612–620.

Yeung, D. S., Cloete, I., Shi, D., & Ng, W. W. Y. (2010). *Sensitivity analysis for neural networks*. Berlin Heidelberg: Springer-Verlag.

Zurada, J. M., Malinowski, A., & Cloete, I. (1994). Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proceedings of IEEE international symposium on circuits and systems - ISCAS '94, vol. 6* (pp. 447–450).